Causal Inference from Uncertain Data

Samantha Kleinberg
Stevens Institute of Technology
CHF progression

Stage A
High Risk

Stage B
Asymptomatic heart failure

Stage C
Symptomatic heart failure

Stage D
Endstage heart failure

Patients have:

- Hypertension
- Coronary artery disease
- Diabetes

- Prior heart attack
- LV dysfunction
- Asymptomatic valvular disease

- Structural heart disease
- Shortness of breath and fatigue
- Reduced exercise tolerance

- Recurrent hospitalization
- Marked symptoms at rest despite intervention

Treatments are:

- Treat high blood pressure
- Encourage exercise

Treatments for A plus, some cases
- ACE inhibitors
- Beta blockers

Treatments for A plus:
- Salt restriction
- ACE inhibitor
- Beta blocker

Treatments for A, B, C and:
- Heart transplant
- Hospice

Adapted from ACC/AHA Guidelines for the Evaluation and Management of Chronic Heart Failure in the Adult Hunt et al. 2001
Electronic health record (EHR) data

• Potential
  – Find drug side effects
  – Markers for disease

• Challenges
  – Cannot do all possible tests
  – Gaps in treatment, change of providers
  – Ambiguity in timing
  – Observation of symptoms, not disease itself
Geisinger CHF data

• 3,838 cases, 28,843 controls
  – Average of 7 years of data, 86 encounters (2.2 years pre diagnosis)

• Variables
  – Comorbidities and problem list
  – Vitals (BP, height, weight, pulse, temperature)
  – Medications
  – Lab values
Causes from EHR data

S. Kleinberg and N. Elhadad (2013) Lessons Learned in Replicating Data-Driven Experiments in Multiple Medical Systems and Patient Populations. AMIA Annual Symposium.
Replication

• Geisinger (~32K patients) in PA
  – Stable, rural population
  – 95% white

• Columbia (~13K patients in AIM clinic) in NY
  – Significant in/out-migration, urban
  – 57% white, 28% black, 15% asian

S. Kleinberg and N. Elhadad (2013) Lessons Learned in Replicating Data-Driven Experiments in Multiple Medical Systems and Patient Populations. AMIA Annual Symposium.
Replication

S. Kleinberg and N. Elhadad (2013) Lessons Learned in Replicating Data-Driven Experiments in Multiple Medical Systems and Patient Populations. AMIA Annual Symposium.
Result comparison

• Some common findings
  – Diabetes (insulin)
  – Thyroid dysfunction (hypothyroidism, thyroid hormone prescription)
  – Hypertension (antihypertensive combination medications, calcium channel/beta blockers)

• Geisinger onset timings likely more accurate, population larger
Uncertain data

- Error in measurements
- Missing data
- Delay
- Inconsistent timescales
Adding uncertainty

\[ P(e|c, x) = \frac{\sum_t ecx}{\sum_t cx} \]

\[ P(e|c, x) = \frac{\sum_t P(e, c, x)}{\sum_t P(c, x)} \]

Causes of changes in glucose

Cohort: 17 subjects with T1DM

Sensor data (collected for >72 hours)

- Glucose values
- Insulin dosage
- Activity
- Sleep stage
- Heart rate
- Temperature

With N. Heintzman (UCSD) 
http://dial.ucsd.edu/what-we-do.php
Results

very vigorous exercise leads to hyperglycemia (fdr < .01) in 5-30 minutes

- Found using both HR (anaerobic activity zone) and METs

• Not found with strict discretization

• Supported by literature
  (Marliss and Vranic, 2002; Riddell and Perkins, 2006)

S. Kleinberg and N. Heintzman. Inference of Causal Relationships from Uncertain Data and Application to Type 1 Diabetes. (under review)
Big data, big problems

- Big ≠ good
- Uncertainty
- Selection bias
- Signal:noise
- Interpretation
- Time
- Ground truth
Future needs

- Hidden variables
- Relationships that change over time
- Simulation