DESIGN CONSIDERATIONS FOR RUNNING HEALTH SYSTEM BASED TRIALS THROUGH THE ELECTRONIC HEALTH RECORD

Benjamin A. Goldstein, PhD, MPH
ben.goldstein@duke.edu

Department of Biostatistics & Bioinformatics
School of Medicine
Duke University

April 17th, 2019
WHY WE WANT TO USE EHRs FOR CLINICAL RESEARCH

- Data readily available
- Often 100,000’s of Patients
- Information collected over a variety of fields
- Can study just about any clinical outcome
- Representative Population
**Why We May *Not* Want to Use EHRs for Clinical Research**

Data are not collected for research

- Data exist in disparate places
- All patients have different pieces of information
- Observational Data
Four Ways EHR Data Differ from Traditional Clinical Data

1. We don’t have everything we want
2. Outcomes are not defined - need to phenotype data
3. Data are both longitudinal and cross-sectional
4. Data not observed randomly - Informed Presence
**Challenge 1:**

*We don’t have everything we want*

- Patients may seek care at multiple facilities
- Most social health information is not recorded or reliable
- Cannot expect death is reliably captured
  - Most people don’t die in the hospital
  - Preliminary work suggests EHRs have only 20% sensitivity
Define local patient population
- Live in the catchment of the health system
- Require a certain number of primary care appointments before eligible for study

Contextual and proxy information can be linked in
- Neighborhood for SES
- Claims data for additional encounters
- NDI/SSDI for death
Challenge 2

Issues of Data Definition: What is a Diabetic?

A comparison of phenotype definitions for diabetes mellitus

Rachel L Richesson,1 Shelley A Rusincovitch,2 Douglas Wixted,2 Bryan C Batch,4 Mark N Feinglos,3 Marie Lynn Miranda,1 W Ed Hammond,2,6 Robert M Califf,2,7 Susan E Spratt8

ABSTRACT

Objective: This study compares the yield and characteristics of diabetes cohorts identified using heterogeneous phenotype definitions.

Materials and methods: Inclusion criteria from seven diabetes phenotype definitions were translated into query algorithms and applied to a population (n=173,503) of adult patients from Duke University Health System. The numbers of patients meeting criteria for each definition and component (glycosuria, diabetes-associated medications, and laboratory results) were compared.

Results: Three phenotype definitions based heavily on ICD-9-CM codes identified 9-11% of the patient population. A broad definition for the Durham Diabetes Coalition included additional criteria and identified 13%. The electronic medical records and genomics, NYC A1c Registry, and diabetes-associated medications definitions, which have restricted or no ICD-9-CM criteria, identified the smallest proportions of patients (7%). The demographic characteristics for all seven phenotype definitions were similar (56-57% women, mean age 60-65 years). Patients in the diabetes-associated medications definition had higher proportions of specific comorbidities and a higher comorbidity burden.

Conclusions: Standardization of diabetes phenotype definitions can streamline the development of healthcare registries and clinical outcomes research. Future work needs to evaluate the impact of these definitions on the interpretation of data from registries and clinical studies.

Table 1: Data domain criteria used in selected phenotype definitions

Table 2: Presence of ICD-9-CM Codes Indicative of Diabetes (n=18,980)

Figure 1: Overlap of diabetes cohorts identified from different categories of phenotype eligibility criteria; n=24,520 patients identified by criteria from any of the three categories.
**Issues of Data Definition: What is a Diabetic?**

<table>
<thead>
<tr>
<th></th>
<th>ICD-9 250.xx</th>
<th>ICD-9 250.x0 &amp; 250.x2 (exclude type I)</th>
<th>Expand. ICD-9 (249.xx, 357.2, 362.0x, 366.41)</th>
<th>HbA1c</th>
<th>Glucose</th>
<th>Abnormal OGTT</th>
<th>Diabetes Meds</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9 250.xx</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMS CCW</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>NYC A1c Registry</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Meds</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>DDC</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>SUPREME-DM</td>
<td>X*</td>
<td>X*</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>eMERGE</td>
<td>X*</td>
<td>X*</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

* Distinction between Inpatient and Outpatient Visits
Definition Differences

Diabetes Validation Results faceted by Endpoint

<table>
<thead>
<tr>
<th>Authoritative Source</th>
<th>ANY</th>
<th>TYPE2</th>
<th>TYPE2unsp</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCW</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDC4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NW</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1C_OR_MED</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity (TPF) vs. 1-Specificity (FPF)
IMPACT OF POORER DEFINITIONS

Bias in Odds Ratio

Sensitivity

Specificity

Odds Ratio

10 / 32
**Challenge 3:**
**Data are both longitudinal and cross-sectional**

- EHR Data consist of *cross-section of longitudinal data*
  - Most data are stored in datamarts that cover fixed periods of time
- Need to use methods for longitudinal data to model updating exposures
  - We most often use time-varying Cox Models
  - Most analyses don’t account for a patient’s trajectory - just most recent value
- Since data are a cross-section no notion of time 0
  - Define “burn-in” periods to define eligibility
  - Use “burn-out” periods to define censoring
Challenge 4
Data are informatively observed: Informed Presence

- Collection of biases due to the fact that patients do not interact randomly with a health system
- Focus on what data are observed as opposed to what are missing
THREE TYPES OF INFORMED PRESENCE

1. We know more about sicker patients
2. Where a patient seeks care is informative
3. Health status driving encounters
## Informed Presence I: Need to Account for Number of Encounters

Regression of Depression on Weight Loss

<table>
<thead>
<tr>
<th></th>
<th>Odds Ratio</th>
<th>$\Delta \log(\text{OR})$</th>
<th>$\Delta \text{OR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimally Adjusted</td>
<td>3.98 (3.81, 4.17)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>+ No. Encounters</td>
<td>2.37 (2.26, 2.50)</td>
<td>-0.52</td>
<td>-1.61</td>
</tr>
<tr>
<td>+ Comorbidities</td>
<td>2.82 (2.69, 2.96)</td>
<td>-0.35</td>
<td>-1.16</td>
</tr>
<tr>
<td>+ No. Encounters &amp; Comorb</td>
<td>2.30 (2.18, 2.42)</td>
<td>-0.55</td>
<td>-1.68</td>
</tr>
</tbody>
</table>
Number of Encounters Potential Confounder

- Weight Loss
- Depr
- Obs Weight Loss
- Obs Depr.
- # Visits

The diagram illustrates the relationships between weight loss, depression, observed weight loss, observed depression, and the number of visits.
## Need to Account for Number of Encounters

<table>
<thead>
<tr>
<th>Condition</th>
<th>Sensitivity</th>
<th>Median Number of Encounters Without Condition</th>
<th>Median Number of Encounters With Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression</td>
<td>56.3%</td>
<td>6</td>
<td>38</td>
</tr>
<tr>
<td>Weight Loss</td>
<td>9.3%</td>
<td>7</td>
<td>45</td>
</tr>
</tbody>
</table>
INFORMED PRESENCE II:
WHERE A PERSON SEeks CARE IS INFORMATIVE

Mean Systolic Blood Pressure

<table>
<thead>
<tr>
<th>SBP</th>
<th>Community</th>
<th>ED</th>
<th>Inpatient</th>
<th>Outpatient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>130</td>
<td>135</td>
<td>140</td>
<td>145</td>
</tr>
</tbody>
</table>
WHERE A PERSON SEeks CARE IS INFORMATIVE

Mean Hemoglobin A1C

ED     Inpatient     Outpatient

- A1C vs. Hemoglobin A1C levels for different healthcare settings.
Location Impacts Inference

- Hazard Ratio for HgB A1C for time to Myocardial Infarction

<table>
<thead>
<tr>
<th>Type</th>
<th>Hazard Ratio</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted</td>
<td>1.06 (1.01, 1.11)</td>
<td>0.026</td>
</tr>
<tr>
<td>Adjusted for Location</td>
<td>0.97 (0.92, 1.02)</td>
<td>0.178</td>
</tr>
<tr>
<td>OP Only</td>
<td>1.07 (1.00, 1.14)</td>
<td>0.044</td>
</tr>
<tr>
<td>ED Only</td>
<td>0.94 (0.89, 0.99)</td>
<td>0.022</td>
</tr>
</tbody>
</table>

- Interaction between A1C and location
WHICH HOSPITAL A PATIENT USES IS INFORMATIVE

Diabetes
N=2,783

Cancer
N=477
### Facility Impacts Inference

#### Odds Ratio for Cancer Status on Diabetes

<table>
<thead>
<tr>
<th>Location</th>
<th>Odds Ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Facilities</td>
<td>1.69</td>
<td>(1.36, 2.10)</td>
</tr>
<tr>
<td>DUMC Only</td>
<td>1.46</td>
<td>(1.15, 1.87)</td>
</tr>
<tr>
<td>DRH Only</td>
<td>0.89</td>
<td>(0.63, 1.26)</td>
</tr>
<tr>
<td>LCHC Only</td>
<td>1.08</td>
<td>(0.74, 1.56)</td>
</tr>
</tbody>
</table>
Referral Hospitals are an Admixed Population

All Cause Survival

Years to Last Known Alive

<table>
<thead>
<tr>
<th>Years</th>
<th>Referral</th>
<th>Non-Referral</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>5522</td>
<td>2114</td>
</tr>
<tr>
<td>0.5</td>
<td>3307</td>
<td>1532</td>
</tr>
<tr>
<td>1.0</td>
<td>2690</td>
<td>1318</td>
</tr>
<tr>
<td>1.5</td>
<td>2158</td>
<td>1110</td>
</tr>
<tr>
<td>2.0</td>
<td>1748</td>
<td>882</td>
</tr>
<tr>
<td>2.5</td>
<td>1360</td>
<td>697</td>
</tr>
<tr>
<td>3.0</td>
<td>995</td>
<td>519</td>
</tr>
<tr>
<td>3.5</td>
<td>637</td>
<td>387</td>
</tr>
<tr>
<td>4.0</td>
<td>474</td>
<td>262</td>
</tr>
<tr>
<td>4.5</td>
<td>282</td>
<td>171</td>
</tr>
<tr>
<td>5.0</td>
<td>65</td>
<td>64</td>
</tr>
</tbody>
</table>
**Admixture Bias**

Comparison of Local and Referral Patients at Cardiac Catheterization Lab

<table>
<thead>
<tr>
<th>Local Patients</th>
<th>Referral Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Older</td>
<td>Younger</td>
</tr>
<tr>
<td>More Comorbidities</td>
<td>More severe valve disease</td>
</tr>
<tr>
<td>Disease due to ageing</td>
<td>Disease due systematic factors</td>
</tr>
<tr>
<td>Better outcomes</td>
<td>More follow-up procedures</td>
</tr>
</tbody>
</table>
INFORMED PRESENCE III: HEALTH STATUS DRIVING ENCOUNTERS

Biomarker Value / Health Status

Show to Clinic

Observed Biomarker Value

Outcome
**Impact of Informative Visit Process on Bias**

**Histogram of Simulated Betas for Biomarker**

(a) All Data Observed

(b) Scheduled Visits Only

(c) Informative Visits

- Frequency
- BIAS

---

*Truth*  
*Simulated Bias*
NEED AN UNDERLYING ASSOCIATION TO INDUCE BIAS

(a) 0% of Population with Scheduled Visits
BM Association with Event
BM Association with IP

(b) 10% of Population with Scheduled Visits
BM Association with Event
BM Association with IP

(c) 30% of Population with Scheduled Visits
BM Association with Event
BM Association with IP

(d) 50% of Population with Scheduled Visits
BM Association with Event
BM Association with IP

(e) 70% of Population with Scheduled Visits
BM Association with Event
BM Association with IP

(f) 100% of Population with Scheduled Visits
BM Association with Event
BM Association with IP
ACCOUNTING FOR NUMBER OF ENCOUNTERS ATTENUATES BIAS

(a) 0% of Population with Scheduled Visits

(b) 10% of Population with Scheduled Visits

(c) 30% of Population with Scheduled Visits

(d) 50% of Population with Scheduled Visits

(e) 70% of Population with Scheduled Visits

(f) 100% of Population with Scheduled Visits
Take Home

- Most analytic challenges arise based on how individuals seek care
- Need to be mindful of what may not be observed in EHR data
- Many challenges are controllable via the study & cohort design
REFERENCES


COLLABORATORS

- Nrupen Bhavsar
- Matt Phelan
- Sarah Peskoе
- Neha Pagidipati