OPPORTUNITIES AND CHALLENGES FOR CLINICAL RESEARCH WITH ELECTRONIC HEALTH RECORDS

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

Department of Biostatistics & Bioinformatics
School of Medicine
Duke University

July 30th, 2017
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
WHAT WE CAN DO WITH ELECTRONIC HEALTH RECORDS

1. Risk Prediction
   - Near term prediction - Risk of inhospital sepsis
   - Long(er) term risk - 30 Day Revisit

2. Population Health
   - Health Service Utilization - Assessment of high utilizers
   - Disease Epidemiology - Experience of incident diabetes in Durham County

3. Comparative Effectiveness Research (CER)
   - Retrospective Studies - Assessment of community intervention for diabetics (SEDI)
   - Prospective Studies - Point of care randomization

4. Association Analyses
   - Risk factors for disease - Phenome Wide Association Studies
   - Data mining - Drug-Drug interactions
**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
1 Structure of Electronic Health Records

2 Research with EHR Data

3 Concluding Thoughts
1 Structure of Electronic Health Records

2 Research with EHR Data

3 Concluding Thoughts
THE EHR FRONT END: GETTING DATA IN
DATA MOVE FRONT END TO DATA WAREHOUSE

- Patient Demographics
- Encounters (Outpatient/Inpatient)
- Diagnoses
- Procedures
- Lab Results
- Medications
- Vital Signs
- Social History
- Radiological Results
- Clinician Notes
- Etc.
It’s Complicated...
Check the Blind Spots

- Data movement and curation requires decision-making.
- Decisions may not be easily accessible.
- Decisions may not be documented or documentation may not be made available.
TURNING EHRs INTO DATA

The analysis pipeline and data platform

Source Data
Examples:
- Enterprise Data Warehouse (EDW)
- External EHR sources
- Electronic data capture systems such as REDCap and eCOS
- Auxiliary data sources such as Census data

Datamart
Core Tables
“Building block data” close to native source data format (often transaction level)

Data Dictionary
Definitions of the source data, and mappings between source and target tables

Curation Dictionary
Datamart-specific processing rules, logic, and algorithms used to create the derived data

Derived Tables
| Aggregations /Summary Levels (eg, summary per year) | Consistency Enforced (eg, excluding adult height of 6 inches) |
| Processing Rules (eg, patient matching and linkage between sources) | Derived Variables (eg, computable phenotypes) |

Analysis and Evaluation
Analytic Dataset Collections
Extracted in output format compatible with statistical purposes, such as SAS. Each collection is structured for the specific analysis and its independent/dependent variables. May include limited or anonymized datasets.

Operational Reporting
Examples: Dashboards and other Business Intelligence (BI) platforms; includes data quality reporting

Figure 2. Datamart components and relationship to external systems and processes.
**Data Marts:**

**Strengths and Weaknesses**

**Strengths**
- Registry like
- Multiple clinical subject areas for cohort
- Regularly scheduled data refresh

**Ideal For:** Posing variety of questions across subject area

**Soft Spots**
- More time and effort to create than data extract
- Structure not easily adaptable
- Data are fixed between refreshes

**Not Ideal For:** Small, targeted analyses
Adding Information Back into EHR

- Dashboards
- Best Practice Alerts
- Predictive Analytics
- Clinical trial recruitment (Snifters)
DIFFERENT TYPES OF CLINIC ENVIRONMENTS

- **Clinic Based System** (e.g. Practice Fusion, Flatiron)
  - Capture Routine Care
  - Local Population
  - Misses inpatient activity

- **Hospital Based System**
  - Observe inpatient procedures and events
  - Only observe when sick
  - Referral hospitals may not represent local or stable population

- **Comprehensive Medical System** (e.g. VA, Kaiser)
  - Observe all types of patient encounters
  - May represent artificial population
1 Structure of Electronic Health Records

2 Research with EHR Data

3 Concluding Thoughts
Collaborative Clinical Research

Biostatistician

Analysis

Study Design

Data Manipulation

Clinical Research

Data Extraction

Variable Definition

Research Question

Epidemiologist/ Clinician

Informaticist
**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 irregularly and potentially densely observed
4. Data not observed randomly - Informed Presence
Most EHRs are Incomplete

- Patients seek care at multiple facilities
- Missing information on when individuals are healthy
- EHRs don’t always contain all the data you want
Linking EHR Data

- Data from other facilities (PCORNet)
- Claims: Center for Medicare & Medicaid Services (CMS)
- Mortality: National Death Index (NDI) & Social Security Death Index (SSDI)
- Genetic Data
- GeoCode Information: American Community Survey (ACS)
- Personal Tracking Data: FitBit, sensors
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, 3 Bryan C Batch, 4 Mark N Feinglos, 4 Marie Lynn Miranda, 1 W Ed Hammond, 4, 5 Robert M Califf, 3, 4
Susan E Spratt 6

OBJECTIVE

This study compared 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 (diagnosis, diabetes-associated medications, and laboratory results) were compared.

RESULTS

Three phenotype definitions based heavily on ICD-9-CM codes identified 9.9-11.1% of the patient population. A broad definition for the Durham Diabetes Coalition included additional criteria and identified 13.0%. The electronic medical records and genomics, NYC, AIC 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 of 72-73 years, and >20% from minority populations). Furthermore, standardization can streamline the development of registries from healthcare data, and inclusion criteria to support regional and national understanding of the populations defined by different phenotype definitions will allow methods for identifying diabetes in large, real-world datasets with uniform sampling criteria. Comparative and aggregate analyses study presents and compares the specifics of patient populations receiving diabetes care, and phenotype definitions adopted from electronic health records and research evidence support intervention programs developed centrally and for federal reporting standards.

BACKGROUND AND SIGNIFICANCE

Diabetes diagnosis and management of diabetes is a complex disease with many factors associated with different criteria.

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></td>
</tr>
<tr>
<td>NYC A1c Registry</td>
<td>X*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

* Distinction between Inpatient and Outpatient Visits
**Definition Differences**

Diabetes Validation Results faceted by Endpoint

![Graph showing sensitivity and specificity for different endpoints.](image)

**Authoritative Source**
- 250
- A1C
- CCW
- DDC4
- MED
- NW
- SUP
- A1C_OR_MED
IMPACT OF POORER DEFINITIONS

Bias in Odds Ratio

Sensitivity

Specificity

Odds Ratio
**Additional Phenotyping Challenges**

- **Death:** Internal work estimates 20% capture of deaths
- **Disease Incidence:** Need to apply ‘burn-in’ periods
- **Censoring:** Need to apply ‘burn-out’ periods
**Multiple Measurements per Person**

**Opportunities**
- Get to observe patient’s evolving health status
- More frequent visits than a typical longitudinal study
- Denser visit information

**Challenges**
- Visits are irregularly spaced
- Different ways to aggregate
- Don’t know what you are not seeing
Look at Changes over Long Periods of Time...
...Or Short Periods of Time

Individual Blood Pressure Curves

Elapsed Time (hrs)

Systolic BP

0 1 2 3
## Analyzing Repeated Measures

<table>
<thead>
<tr>
<th>Summarizing Data</th>
<th>Modelling Progression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean/Median Values</td>
<td>Regression Splines</td>
</tr>
<tr>
<td>Extreme Values</td>
<td>Functional Data Analysis</td>
</tr>
<tr>
<td>Variability</td>
<td>Joint Models</td>
</tr>
<tr>
<td>Number of Measurements</td>
<td></td>
</tr>
</tbody>
</table>
Simpler Methods Often Work Best
EHR DATA OPTIMIZED FOR NEARER TERM PREDICTION

ROC Curves for Forecasting SCD

Days Out:
- 1 Day (c-stat= 0.803)
- 30 Days (c-stat= 0.729)
- 90 Days (c-stat= 0.714)
- 180 Days (c-stat= 0.685)
- 365 Days (c-stat= 0.642)
## Top Predictors

<table>
<thead>
<tr>
<th>1 Day</th>
<th>7 Days</th>
<th>30 Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 LabValue: Albumin</td>
<td>LabValue: Albumin</td>
<td>LabValue: Albumin</td>
</tr>
<tr>
<td>2 Pre Systolic BP</td>
<td>Pre Systolic BP</td>
<td>Pre Systolic BP</td>
</tr>
<tr>
<td>3 Pre MAP</td>
<td>Pre MAP</td>
<td>Lowest Systolic BP</td>
</tr>
<tr>
<td>4 Pre Pulse Pressure</td>
<td>LabValue: WBC</td>
<td>LabValue: Creatinine</td>
</tr>
<tr>
<td>5 LabValue: Hemoglobin</td>
<td>Medication Dose: Epogen</td>
<td>Pre MAP</td>
</tr>
<tr>
<td>6 Lowest Systolic BP</td>
<td>LabValue: Creatinine</td>
<td>Post MAP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>90 Days</th>
<th>180 Days</th>
<th>365 Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 LabValue: Albumin</td>
<td>LabValue: Albumin</td>
<td>LabValue: Albumin</td>
</tr>
<tr>
<td>2 Pre Weight</td>
<td>Pre Weight</td>
<td>Medication Dose: Epogen</td>
</tr>
<tr>
<td>3 Pre Systolic BP</td>
<td>Pre Map</td>
<td>Age</td>
</tr>
<tr>
<td>4 Pre Pulse Pressure</td>
<td>Post Weight</td>
<td>LabValue: Creatinine</td>
</tr>
<tr>
<td>5 Medication Dose: Epogen</td>
<td>Medication Dose: Epogen</td>
<td>Pre Systolic BP</td>
</tr>
<tr>
<td>6 Post Weight</td>
<td>Pre Systolic BP</td>
<td>Pre Pulse Pressure</td>
</tr>
</tbody>
</table>
## Top Predictors

<table>
<thead>
<tr>
<th>1 Day</th>
<th>7 Days</th>
<th>30 Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>LabValue: Albumin</td>
<td>LabValue: Albumin</td>
<td>LabValue: Albumin</td>
</tr>
<tr>
<td>Pre Systolic BP</td>
<td>Pre Systolic BP</td>
<td>Pre Systolic BP</td>
</tr>
<tr>
<td>Pre MAP</td>
<td>Pre MAP</td>
<td>Lowest Systolic BP</td>
</tr>
<tr>
<td>Pre Pulse Pressure</td>
<td>LabValue: WBC</td>
<td>LabValue: Creatinine</td>
</tr>
<tr>
<td>LabValue: Hemoglobin</td>
<td>Medication Dose: Epogen</td>
<td>Pre MAP</td>
</tr>
<tr>
<td>Lowest Systolic BP</td>
<td>LabValue: Creatinine</td>
<td>Post MAP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>90 Days</th>
<th>180 Days</th>
<th>365 Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>LabValue: Albumin</td>
<td>LabValue: Albumin</td>
<td>LabValue: Albumin</td>
</tr>
<tr>
<td>Pre Weight</td>
<td>Pre Weight</td>
<td>Medication Dose: Epogen</td>
</tr>
<tr>
<td>Pre Systolic BP</td>
<td>Pre Map</td>
<td>Age</td>
</tr>
<tr>
<td>Pre Pulse Pressure</td>
<td>Post Weight</td>
<td>LabValue: Creatinine</td>
</tr>
<tr>
<td>Medication Dose: Epogen</td>
<td>Medication Dose: Epogen</td>
<td>Pre Systolic BP</td>
</tr>
<tr>
<td>Post Weight</td>
<td>Pre Systolic BP</td>
<td>Pre Pulse Pressure</td>
</tr>
</tbody>
</table>
Different Data Elements have Different Predictability

Predicting Death at Different Horizons
With Different Data Sources

- Demographics (n = 5)
- Service Utilization (n = 4)
- Comorbidities (n = 48)
- Meds (n = 30)
- Labs (n = 21)
- Vitals (n = 12)
- All (n = 130)

Time Horizon:
- 7d
- 30d
- 90d
- 180d
- 1yr
- 2yr
- 3yr

C-Statistics
Biases in EHRs: Informed Presence

- We only see patients when they are sick
- We only see information that is deemed important
- Different environments have different policies
INFORMED PRESENCE I: WHERE A PERSON SEeks CARE IS INFORMATIVE

Mean Hemoglobin A1C

- ED
- Inpatient
- Outpatient
**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
INFORMED PRESENCE II:
 WHICH HOSPITAL A PATIENT USES IS INFORMATIVE

![Diabetes Venn Diagram](image1)

- **Diabetes**: N=2,783
- **LCHC**: 6792
- **DUMC**: 381
- **DRH**: 126
- **456**:
- **941**:
- **225**:

![Cancer Venn Diagram](image2)

- **Cancer**: N=477
- **LCHC**: 9091
- **DUMC**: 205
- **DRH**: 39
- **76**:
- **47**:
- **14**:
- **20**:

### Facility Impacts Inference

<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>

Odds Ratio for Cancer Status on Diabetes
INFORMED PRESENCE III:

REFERAL HOSPITALS ARE AN Admixed POPULATION
### Admixture Bias

Comparison of Local and Referal Patients at Cardiac Catheterization Lab

<table>
<thead>
<tr>
<th>Local Patients</th>
<th>Referal 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>

Disease due to ageing vs. Disease due systematic factors.
**INFORMED PRESENCE IV: NEED TO ACCOUNT FOR NUMBER OF ENCOUNTERS**

Regression of Depression on Weight Loss

<table>
<thead>
<tr>
<th></th>
<th>Odds Ratio</th>
<th>Δ log(OR)</th>
<th>Δ 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
- # Visits
- Obs Weight Loss
- Obs Depr.
**NEED TO ACCOUNT FOR NUMBER OF ENCOUNTERS**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Sensitivity</th>
<th>Median Number of Encounters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression</td>
<td>56.3%</td>
<td>6</td>
</tr>
<tr>
<td>Weight Loss</td>
<td>9.3%</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>With Condition:</td>
</tr>
<tr>
<td>Depression</td>
<td></td>
<td>38</td>
</tr>
<tr>
<td>Weight Loss</td>
<td></td>
<td>45</td>
</tr>
</tbody>
</table>
NUMBER OF ENCOUNTERS POTENTIAL CONFOUNDER

Bias In Estimated Association

![Graph showing the bias in estimated association with and without adjustment. The x-axis represents the probability of capturing exposure, ranging from 0.2 to 1.0, and the y-axis represents the bias ranging from -0.5 to 1.0. The graph compares 'No Adjustment' (purple line) and 'Adjustment' (green line).]
1. Structure of Electronic Health Records

2. Research with EHR Data

3. Concluding Thoughts
**Extra Care Needed**

- Need to be mindful from where the data come
- There is not always one way to turn raw data into analytic data
- Which data to *cut* is more important than how you analyze it
- New analytic techniques may be useful/necessary
**Questions to ask when designing EHR based studies**

- Where in the health system are the data collected?
- What is the coverage/catchment area of your health system?
- Is the patient population receiving care across multiple institutions/centers?
- Do the data constitute different catchments? (Admixture)
- How are you defining exposures and outcomes? (Phenotyping)
- How are you defining person-time?
  - What is an appropriate burn-in period to define a cohort?
  - Is a burn-out period necessary to define censoring?
- Do different populations produce more information (i.e. sicker patients have more encounters)?
**Additional Frontiers**

- Micro-randomized trials
- Integration of external data
- Real time risk assessment
### Is it all bad?

#### A lot of opportunities with EHRs

- More studies
- Cheaper studies
- Faster studies
- (Perhaps) More representative studies
REFERENCES


