Practical Methods for Working with EHR Data

CIDA Big Data Seminar Series

Laura K. Wiley, PhD
Assistant Professor
Biomedical Informatics & Personalized Medicine

@GenomeGal

University of Colorado
Anschutz Medical Campus

COLORADO CENTER FOR
PERSONALIZED MEDICINE
Overview

• Introduction to EHRs
• Landscape of EHR-based Research
  • Types & Limitations of EHR Data
  • Phenome-Wide Association Studies
  • Computational Phenotyping
• CU Resources and How to Learn More
What is an EHR?
How to Leverage EHRs for Research?

Clinical Trial Recruitment

Analytics/ Learning Healthcare System

Population Health/ Epidemiology

Link to Biobank for Genetic Research
EHRs Require Paradigm Shift
EHR – Based Research Landscape

Scope

Focused

Broad

Simple

Complex

Analysis

ML/Predictive Analytics

Computational Phenotyping

PheWAS
What Types of Data are in an EHR?

## Data bases

<table>
<thead>
<tr>
<th>Diagnosis codes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fake ID</strong></td>
</tr>
<tr>
<td>34068</td>
</tr>
<tr>
<td>37650</td>
</tr>
<tr>
<td>14060</td>
</tr>
<tr>
<td>23315</td>
</tr>
<tr>
<td>76296</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lab tests</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fake ID</strong></td>
</tr>
<tr>
<td>3536</td>
</tr>
<tr>
<td>72921</td>
</tr>
<tr>
<td>102460</td>
</tr>
<tr>
<td>135043</td>
</tr>
<tr>
<td>135432</td>
</tr>
</tbody>
</table>

## Structured

### Problem lists:
- **Medications known to be prescribed:**
  - Keptra 750 mg 1/2 tab q am and pm
  - Dexamet 60 mg by mouth daily
  - Aspirin 325 mg 1 tablet by mouth daily
  - Clopidogrel 75 mg tablet 1 tablet by mouth daily
- **Known adverse and allergic drug reactions:**
  - Sulfas Drugs
- **Known significant medical diagnoses:**
  - Seizure disorder
  - Anemia
  - Heartburn
- **Known significant operative and invasive procedures:**
  - 2003 Appendectomy
  - 2005 Stents put in **DATE [Aug 29 05]**

## Semi-Structured

### Clinical notes

**EXAM: BILATERAL DIGITAL SCREENING MAMMOGRAM WITH CAD, **DATE[Mar 16 01]**, COMPARISON, **DATE[Jul 01 01]**

**TECHNIQUE:** Standard CC and MLO views of both breasts were obtained.

**FINDINGS:** The breast parenchyma is heterogeneously dense. The pattern is extremely complex with postsurgical change seen in the right upper outer quadrant and scattered benign appearing calcification seen bilaterally. A possible asymmetry is seen in the superior aspect of the left breast. The parenchymal pattern otherwise remains stable bilaterally with no new distortion or suspicious calcifications. **IMPRESSION:**

**RIGHT:** No interval change. No current evidence of malignancy. **LEFT:** Possible developing asymmetry superior aspect left breast for which further evaluation by true lateral and spot compression views recommended. Ultrasound may also be needed. **RECOMMENDATION:** Internal mammary and additional imaging as outlined above. A left breast ultrasound may also be needed. **BI-RADS Category 6:** Incomplete Assessment - Need additional imaging evaluation. **IMPRESSION:**

**RIGHT:** No interval change. No current evidence of malignancy....

## Unstructured

---

Wei, et. al.

Genome Med 2015

PMID: 25937834
Structured Data – ICD Codes

- International Classification of Disease (ICD)
- In US, we use ICD-10-CM (ICD-9-CM < October 2015)
- Diagnostic codes:
  - ICD-9-CM: ~13,500
  - ICD-10-CM: >70,000

W59.22XD
Struck by Turtle, Subsequent Encounter
Structured Data – ICD Codes

False Positives

False Negatives
Structured Data – Laboratory Data

**Positives**
- Closer to Biology!
- Continuous Variable
  - Moar Power!
  - No controls!
- Look at Extremes

**Challenges**
- Multiple panels w/ same Labs
- Multiple lab values/ person
- Method & Reference range change over time
- UNITS!
- Junk
  - (PHONE, /, <4, Duplicate Order, Mislabeled Specimen, Result)
Structured(ish) Data - Medications

**Formatting**
- E-Prescribed: Name, size, frequency structured
- Dosing Instructions often semi or unstructured
- Also found in problem list and clinical notes -> unstructured

**Challenges**
- Atorvastatin = Lipitor
- Atorvastatin + Ezetimibe = Liptruzet
- Atorvastatin/Simvastatin both HMG-CoA Reductase Inhibitors
- Atorvastatin & Ezetimibe both cholesterol lowering drugs
- Order -> Fill -> Taking the Drug
- Identifying Discontinuation Difficult
Unstructured Data – Clinical Text

• Goal: Convert Unstructured Text -> Structured Data

• NLP Tools:
  • **Concept-indexing:**
    • “mad cow disease” → C0120202
    • “Bovine Spongiform Encephalopathy” → C0120202
  • **Negation/certainty tagging:**
    • “no evidence of diabetes”
  • **Identifying temporal expressions:**
    • “colonoscopy 5 years ago”
  • ... and much more
EHR – Based Research Landscape

Scope

Broad

Focused

Analysis

Simple

Complex

PheWAS

ML/Predictive Analytics

Computational Phenotyping
What is (Computational) Phenotyping?

- Translates observed characteristics or manifestations of a human condition/disease
- Utilizes clinical criteria from the EMR to identify subjects meeting definition
What is Computational Phenotyping?

Definite Cases (algorithm-defined)  Possible Cases (require manual review)  Excluded (algorithm-defined)  Controls (algorithm-defined)
Computational Phenotyping Overview

Population

Inclusion
Must have

Exclusion
Must NOT Have

Case/Control Definitions
Steroid-Induced Osteonecrosis

Population: Vanderbilt BioVU

Inclusion:
Any Corticosteroid, No Minimum Dose
Intravenous, Intramuscular, or Oral (Not Inhaled)
At least 14 days

Exclusion:
Alcohol Abuse, Sickle Cell, Gaucher, Legge-Calve-Perthes, HIV, Organ Transplant

Cases:
ICD9: 733.4X
OR
Keywords: osteonecrosis, avascular necrosis, osteochondritis dessicans

Controls:
Not a Case
AND
No record of bisphosphonates
How Do You Know It Worked?

Gold Standard Population

Data

Manually Review All or Random Subset of Records

Gold Standard Population

Algorithm

Gold Standard Truth

+ True Positive False Negative

- False Positive True Negative
Caution – Danger Ahead

- Defining Controls is HARD
  - Clinical?
  - Biological?
EHR – Based Research Landscape

Scope

Focused

Simple

Analysis

Complex

ML/Predictive Analytics

Computational Phenotyping

PheWAS
Phenome-Wide Association Study (PheWAS)

When to Use PheWAS?
How Does PheWAS Work?

ICD billing code in EMR → PheWAS code

- **556.0** Ulcerative (chronic) enterocolitis
- **556.5** Left-sided ulcerative (chronic) colitis
- **556.9** Ulcerative colitis, unspecified
- **648.21** Anemia of mother, with delivery
- **648.23** Anemia, antepartum

- **555.2** Ulcerative colitis
- **555.2** Anemia associated with pregnancy
How Does PheWAS Work?

Does individual have any incidence of ICD9s that map to PheWAS code?

Yes → Does individual have 2+ incidences of ICD9s on different days that map to PheWAS code?

Yes → Case for PheWAS code

No → Does individual have any exclusions for PheWAS code?

Yes → Exclusion for PheWAS code (NA)

No → Control for PheWAS code

Repeat for each PheWAS code and Perform Regressions
PheWAS R Package

- [https://github.com/PheWAS/PheWAS](https://github.com/PheWAS/PheWAS)
- Contains:
  - Translation Tables (ICD9-CM, ICD10-CM, ICD10)
  - Analysis Scripts
  - Multiple Testing Correction (Bonferroni & FDR)
You Have A Significant Hit! Now What?

- PheWAS Code Correlation – Which is the driver?
- Check Genotype & Phenotype Frequencies
- Is this plausible/meaningful?
- Look at individual patients w/ Code

**PheWAS codes are NOT phenotypes**
EHR – Based Research Landscape

Scope

Focused

Broad

Analysis

Simple

Complex

ML/Predictive Analytics

Computational Phenotyping

PheWAS
CU Resources

- Clinical Trial Recruitment
- Analytics/ Learning Healthcare System
- Population Health/ Epidemiology
- Link to Biobank for Genetic Research

EMR Repository

TriNetX
REDCap
Center for Innovative Design & Analysis
Colorado Center for Personalized Medicine
Biobank
D2V@CU
Interested in Learning More?

Clinical Data Science Specialization
Created By:
University of Colorado
Anschoitz Medical Campus
LearnClinicalDataScience.org

Instructions
1. Log in to my.cu.edu
2. Open “Training” (if student) or “CU Resources Home” > “Training” (if staff/faculty)
3. Select “CU on Coursera”
4. Click “Join for Free”
5. Click “Log in with University of Colorado”

FREE For CU Faculty, Staff, & Students!
How To Get Help

• Health Data Compass: www.healthdatacompass.org
• Colorado Center for Personalized Medicine Biobank: www.cobiobank.org
• CIDA: 
  http://www.ucdenver.edu/academics/colleges/PublicHealth/research/centers/CBC/Pages/welcome.aspx
• D2V: 
  http://www.ucdenver.edu/academics/colleges/medicalschool/programs/D2V/Pages/D2V.aspx