Improving EHR Data Quality with Automated Phenotyping

Shawn Murphy MD, Ph.D.
Partners Healthcare and Harvard Medical School
October 10th, 2019
Integration strategy from “Informatics for Integrating Biology and the Bedside (i2b2)” sponsored by the National Institutes of Health, what is it?

- Software for explicitly organizing and transforming person-oriented clinical data to a way that is optimized for clinical genomics research
  - Allows integration of clinical data, trials data, and genotypic data

- A portable and extensible application framework
  - Software is built in a modular pattern that allows additions without disturbing core parts
  - Available as open source at https://www.i2b2.org
Interrogation can occur through i2b2 web client
Running a Query

The image shows a screenshot of a clinical image bank interface. The interface includes a query tool with sections for navigating terms, query timing, and a non-temporal query. The query is for patients with epilepsy and recurrent seizures, and the result shows that 29 patients meet the criteria.
I2b2 Community Software Modules contributed as “Cells”

https://community.i2b2.org/wiki/display/i2b2/i2b2+Community+Projects
i2b2 Cell: The Canonical Software Module

HTTP XML
(minimum: RESTful)
An i2b2 Environment is built from i2b2 Cells
Federated Queries

- Partners HealthCare System
- Boston Children’s Hospital
- BIDMC
- Boston Health Net (BMC and Community Health Centers)
- Columbia U. Medical Center and New York Presbyterian Hospital
- Wake Forest Baptist Medical Center
- Morehouse/Grady/RCMI
- University of California, Davis
- Washington University in St. Louis
- U Texas Health Science Center/Houston
Implementations

CTSA’s

- Boston University
- Case Western Reserve University (including Cleveland Clinic)
- Children’s National Medical Center (GWU), Washington D.C.
- Duke University
- Emory University (including Morehouse School of Medicine and Georgia Tech)
- Harvard University (including Beth Israel Deaconness Medical Center, Brigham and Women’s Hospital, Children’s Hospital Boston, Dana Farber Cancer Center, Joslin Diabetes Center, Massachusetts General Hospital)
- Medical University of South Carolina
- Medical College of Wisconsin
- Oregon Health & Science University
- Penn State Milton S. Hershey Medical Center
- Tufts University
- University of Alabama at Birmingham
- University of Arkansas for Medical Sciences
- University of California Davis
- University of California, Irvine
- University of California, Los Angeles*
- University of California, San Diego*
- University of California San Francisco
- University of Chicago
- University of Cincinnati (including Cincinnatni Children’s Hospital Medical Center)
- University of Colorado Denver (including Children’s Hospital Colorado)
- University of Florida
- University of Kansas Medical Center
- University of Kentucky Research Foundation
- University of Massachusetts Medical School, Worcester
- University of Michigan
- University of Pennsylvania (including Children’s Hospital of Philadelphia)
- University of Pittsburgh (including their Cancer Institute)
- University of Rochester School of Medicine and Dentistry
- University of Texas Health Sciences Center at Houston
- University of Texas Health Sciences Center at San Antonio
- University of Texas Medical Branch (Galveston)
- University of Texas Southwestern Medical Center at Dallas
- University of Utah
- University of Washington
- University of Wisconsin - Madison (including Marshfield Clinic)
- Virginia Commonwealth University
- Weill Cornell Medical College

Academic Health Centers (does not include AHCs that are part of a CTSA):

- Arizona State University
- City of Hope, Los Angeles
- Georgia Health Sciences University, Augusta
- Hartford Hospital, CN
- HealthShare Montana
- Massachusetts Veterans Epidemiology Research and Information Center (MAVERICK), Boston
- Nemours
- Phoenix Children’s Hospital
- Regenstrief Institute
- Thomas Jefferson University
- University of Connecticut Health Center
- University of Missouri School of Medicine
- University of Tennessee Health Sciences Center
- Wake Forest University Baptist Medical Center

HMOs:

- Group Health Cooperative
- Kaiser Permanente

International:

- Georges Pompidou Hospital, Paris, France
- Hospital of the Free University of Brussels, Belgium
- Inserm U936, Rennes, France
- Institute for Data Technology and Informatics (IDI), NTNU, Norway
- Institute for Molecular Medicine Finland (FIMM)
- Karolinska Institute, Sweden
- Landspitali University Hospital, Reykjavik, Iceland
- Tokyo Medical and Dental University, Japan
- University of Bordeaux Segalen, France
- University of Erlangen-Nuremberg, Germany
- University of Goettingen, Goettingen, Germany
- University of Pavia, Pavia, Italy
- University of Seoul, Seoul, Korea

Companies:

- Johnson and Johnson (TransMART)
- GE Healthcare Clinical Data Services
I2b2 Software components are distributed as open source
Recent Community Project Releases

- Incorporation of FHIR into i2b2
  - FHIR Cell to allow single patient’s data to be returned in FHIR
  - FHIR Ontology and ETL to allow direct import of FHIR into i2b2
  - FHIR Cell to allow i2b2 to extend to FHIR query endpoints

- i2b2-based system for Accruing Patients for Clinical Trials
  - Extensions to SHRINE for management of clinical trials
  - Web Client Plug-ins to extend SHRINE queries to local i2b2-based patient recruitment

- i2b2-based system to query and return data from Observation-Fact tables contained in multiple different i2b2 Hives
Improve Quality of i2b2 Queries through Machine Learning
Use Phenotyping Algorithms to define cohorts of treatment-resistant and treatment-responsive depression.

Initially: AUC = 0.54
Finally: AUC = 0.87

Need to Determine: Depressed or Well at Encounter

Must Improve Accuracy of Diagnoses from Electronic Health Record

<table>
<thead>
<tr>
<th>Clinical Status</th>
<th>Model</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Precision</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depressed</td>
<td>Billing Codes</td>
<td>0.95</td>
<td>0.09 (0.03)</td>
<td>0.57 (0.14)</td>
<td>0.54 (0.02)</td>
</tr>
<tr>
<td>Depressed</td>
<td>NLP</td>
<td>0.95</td>
<td>0.42 (0.05)</td>
<td>0.78 (0.02)</td>
<td>0.88 (0.02)</td>
</tr>
<tr>
<td>Depressed</td>
<td>NLP + Billing Codes</td>
<td>0.95</td>
<td>0.39 (0.06)</td>
<td>0.78 (0.02)</td>
<td>0.87 (0.02)</td>
</tr>
<tr>
<td>Well</td>
<td>Billing Codes</td>
<td>0.95</td>
<td>0.06 (0.02)</td>
<td>0.26 (0.27)</td>
<td>0.55 (0.03)</td>
</tr>
<tr>
<td>Well</td>
<td>NLP</td>
<td>0.95</td>
<td>0.37 (0.06)</td>
<td>0.86 (0.02)</td>
<td>0.85 (0.02)</td>
</tr>
<tr>
<td>Well</td>
<td>NLP + Billing Codes</td>
<td>0.95</td>
<td>0.39 (0.07)</td>
<td>0.85 (0.02)</td>
<td>0.86 (0.02)</td>
</tr>
</tbody>
</table>
Use NLP to extract the relevant features from the set of patient notes.
Data Integration in Biobank Portal

Electronic Medical Record (EMR) Data
- RPDR
  - Coded Data
    - Demographics
    - Diagnoses
    - Lab Results
  - Text Data (Notes/Reports)
    - Medications
    - Procedures
    - Visits
    - Physician Notes
    - Imaging Reports
    - Pathology Reports
    - Surgery Notes

Informatics Tools
- Calculated Controls (Charlson Index)
- Data Visualization
- Data Queries
- Annotation
- Extract Data
- Natural Language Processing

Additional Data
- Other Research Data
- Survey Data

Genetic Data
- GWAS

Biobank Data
- Samples
  - DNA
  - Serum
  - Plasma
- Consent
  - Recontact
  - Consent Status

Validated Phenotypes
- Type II Diabetes
- Coronary Artery Disease
- Congestive Heart Failure
- Multiple Sclerosis
- Rheumatoid Arthritis
- IBD
- Bipolar Disorder

Research
Curating a Disease Algorithm with a Gold Standard

1. **Create a gold standard training set.**

2. **Create a comprehensive list of features** from patient’s electronic data that describe the disease of interest.

3. **Develop the classification algorithm.** Using the data analysis file and the training set from step 1, assess the frequency of each variable. Remove variables with low prevalence. Apply adaptive LASSO penalized logistic regression to identify highly predictive variables for the algorithm.

4. **Apply the algorithm to all subjects** in the superset and assign each subject a probability of having the phenotype.
Curating a Disease Algorithm with a Silver Standard

1. Query for total number of mentions of disease

2. Fit the mentions of disease to two curves normalizing for # of visits

3. Resolve the curves and separate into two groups, these are actually equal to patients with and without the disease

4. Apply the algorithm to all subjects and assign each subject a probability of having the phenotype
<table>
<thead>
<tr>
<th>category</th>
<th>PheWAS_code</th>
<th>abbr</th>
<th>PheWAS_name</th>
<th>model</th>
<th>ICD_PPV</th>
<th>ICD_AUC</th>
<th>AUC</th>
<th>PPV</th>
<th>TPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONC</td>
<td>PheWAS:189.21</td>
<td>BLCA</td>
<td>Bladder cancer</td>
<td>PheNorm_ICD</td>
<td>0.80</td>
<td>0.903</td>
<td>1.00</td>
<td>1.00</td>
<td>0.42</td>
</tr>
<tr>
<td>ONC</td>
<td>PheWAS:204</td>
<td>LEUK</td>
<td>Leukemia</td>
<td>PheNorm_ICD</td>
<td>0.73</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.91</td>
</tr>
<tr>
<td>PSYCH</td>
<td>PheWAS:297.1</td>
<td>SI</td>
<td>Suicidal ideation</td>
<td>PheNorm_ICDNLP</td>
<td>0.93</td>
<td>0.786</td>
<td>1.00</td>
<td>1.00</td>
<td>0.43</td>
</tr>
<tr>
<td>PSYCH</td>
<td>PheWAS:305.2</td>
<td>EATD</td>
<td>Eating disorder</td>
<td>PheNorm_ICDNLP</td>
<td>0.53</td>
<td>0.482</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>NEURO</td>
<td>PheWAS:327.4</td>
<td>INSOM</td>
<td>Insomnia</td>
<td>PheNorm_ICDNLP</td>
<td>0.93</td>
<td>0.821</td>
<td>1.00</td>
<td>1.00</td>
<td>0.50</td>
</tr>
<tr>
<td>CARDIO</td>
<td>PheWAS:452.2</td>
<td>DVT</td>
<td>Deep vein thrombosis</td>
<td>PheNorm_ICDNLP</td>
<td>0.87</td>
<td>0.692</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>NEURO</td>
<td>PheWAS:817</td>
<td>CONC</td>
<td>Concussion</td>
<td>PheNorm_NLP</td>
<td>0.73</td>
<td>0.682</td>
<td>1.00</td>
<td>1.00</td>
<td>0.27</td>
</tr>
<tr>
<td>METAB</td>
<td>PheWAS:250.1</td>
<td>T1DM</td>
<td>Type 1 diabetes</td>
<td>PheNorm_ICD</td>
<td>0.17</td>
<td>0.882</td>
<td>0.984</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>ONC</td>
<td>PheWAS:184.11</td>
<td>OVCA</td>
<td>Ovarian cancer</td>
<td>PheNorm_ICDNLP</td>
<td>0.60</td>
<td>0.926</td>
<td>0.981</td>
<td>1.00</td>
<td>0.67</td>
</tr>
<tr>
<td>ONC</td>
<td>PheWAS:182</td>
<td>UTCA</td>
<td>Uterine cancer</td>
<td>PheNorm_ICD</td>
<td>0.50</td>
<td>0.867</td>
<td>0.980</td>
<td>1.00</td>
<td>0.86</td>
</tr>
<tr>
<td>GI</td>
<td>PheWAS:555.1</td>
<td>CD</td>
<td>Crohn's disease</td>
<td>PheNorm_mean</td>
<td>0.54</td>
<td>0.961</td>
<td>0.980</td>
<td>0.90</td>
<td>0.97</td>
</tr>
</tbody>
</table>
RESULT
Accurate and Simple Disease Labels for Queries

Complicated

Simple
InDigitally DrivenHealthcare, DiseaseLabels determine Algorithms for ManagingPatient

Mary
Followed for:
- Prevention
- Diabetes Type II
- Hypertension
- Heart Failure

APPSTOMANAGE DIET AND EXERCISE

GLUCOSE MONITORING

**Metformin: Hb A1c**

<table>
<thead>
<tr>
<th>Day</th>
<th>Metformin</th>
<th>SGLT-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Readings between 6 and 7 mmol/L indicate a normal range.*
*Readings above 7 mmol/L indicate high risk of DKA. Contact your healthcare team immediately for advice.*
Innovations Enabling Digital Care

Digital and IoT devices continuously output Patient Data.


Navigator Model dramatically increases Frequency and Convenience for Patient Communication.

System drives Pragmatic Clinical Trials Leading to Continuous Process Improvement.
Digital Twin for Continuously Assessing Patient

Data Enclave
Multiple query endpoints feeding into Data Commons for Research, Operations, and Clinical Insights

Computation
Components:
• Cohort Creation and focused data extraction
• Managed Terminology
• Regulatory Oversight
• Libraries of Python/R
• Commonly formatted data

FHIR based Apps
12b2 tranSMART
Education Jupyter Notebook
Python/R Environment
SAS/SPSS MATLAB

Clinical Analytics

Data Enclave
1. Enable Data Extracts, perhaps some are Federated
2. Combine and Link data, put in common OMOP/i2b2 format
3. Conduit to Data Enclave

Technical Solution Development
4. Healthcare Ready Bundle

Research and Clinical Application Projects
5. FHIR based SMART Apps
6. 12b2 tranSMART with Fractalis plugin (next version of SmartR plugin)
7. Jupyter Notebook with AI Visualizations – code can advance to production
8. Python/R Environment full interactive development in Data Lake
Combining i2b2 with tranSMART
i2b2 used for Big Clinical Data

1) Queries for aggregate patient numbers

- Warehouse of in & outpatient clinical data
- 6.7 million Partners Healthcare patients
- 3.1 billion diagnoses, medications, genomics, procedures, laboratories, & physical findings coupled to demographic & visit data
- Authorized use by faculty status
- Clinicians can construct complex queries
- Queries cannot identify individuals, internally can produce identifiers for (2)

2) Returns detailed patient data

- Start with list of specific patients, usually from (1)
- Authorized use by IRB Protocol
- Returns contact and PCP information, demographics, providers, visits, diagnoses, medications, procedures, laboratories, microbiology, reports (discharge, LMR, operative, radiology, pathology, cardiology, pulmonary, endoscopy), and images into a Microsoft Access database and text files.
Enabled by Star Schema

**Concept DIMENSION**
- concept_key
- concept_text
- search_hierarchy

**Patient Concept FACTS**
- patient_key
- concept_key
- start_date
- end_date
- practitioner_key
- encounter_key
- value_type
- numeric_value
- textual_value
- abnormal_flag

**Patient DIMENSION**
- patient_key
- patient_id (encrypted)
- sex
- age
- birth_date
- race
- deceased
- ZIP

**Encounter DIMENSION**
- encounter_key
- encounter_date
- hospital_of_service

**Practitioner DIMENSION**
- practitioner_key
- name
- service

**Patient Concept FACTS**
3100 million

**Encounter DIMENSION**
450

**Practitioner DIMENSION**
6.4
0.04

**Binary Tree**

start
search
Start Schema enables both i2b2 and tranSMART

i2b2

CLINICAL HOSPITAL DATA

REDCap

tranSMART

CLINICAL TRIAL DATA

Fractalis
Risks with i2b2 - tranSMART common Database Instance

- Privacy Models different
- Database collisions from simultaneous actions
  - Principle risk is updating
- Index optimization different for different use cases
- Inherent assumptions about data are different
  - "One fact" per study visit regardless of time
Tribute to…

- Jeff Klann
- Michael Mendis
- Kavi Wagholikar
- Lori Phillips
- Isaac Kohane
- Kenneth Mandl
- Joshua Mandel
- Griffin Weber
- Paul Avillach
- Christopher Herrick
- Vivian Gainer
- Victor Castro
- Nich Wattanasin
- Wayne Chan
- David Wang
- Andrew Cagan
- Bhaswati Ghosh
- Retta Metta
- Adam Landman
- Willian Gordon
I2b2, SHRINE, and SMART Information and Software on the Web

i2b2 Homepage (https://www.i2b2.org)

i2b2 Software (https://www.i2b2.org/software)

i2b2 Community Site (https://community.i2b2.org)

SMART Platforms Homepage (http://smarthealthit.org)

Partners Healthcare, NIH/NCBC/BD2K; /NIMH; /NCATS; /NIBIB; /NHGRI

NIH R01 EB014947
NIH U54 LM008748
NIH U01 HG008685
PCORI 282364.5077585.0007

NIH U54 HG007963
NIH R01 AT006364
NIH R01 AT005280
NIH P01 AT006663
THANK YOU