Applications built on the OHDSI OMOP framework

OHDSI Community Call

Ben Glicksberg, PhD
2/5/19
Butte Lab
Bakar Computational Health Sciences Institute
University of California, San Francisco (UCSF)
Clinical Informatics in the era of big data

“The Quantified Self”

Input:
- Genetic Architecture
- Predisposition
- Environment
- Family History
- Gene x Environment Interaction
- Exposures
- Transcriptome
- Microbiome
- Epigenome
- Proteome
- Dynamics

System:
- Artificial Intelligence
- Machine Learning

Output:
- Intervventional Therapies
- Risk Factor Identification
- Disease Monitoring
- Patient Monitoring
- Therapeutic Stratification

Electronic Health Records

Patient Visit
- Demographics
  - Name
  - Address
  - Telephone
  - Sex
  - Date of Birth
  - Ethnicity
  - Race
  - Religion
  - Insurance

- Chief Complaint
- Family History
- Vital Signs
- Blood Pressure
- Pulse
- Temperature
- Medications
- Allergies
- Disease History
- Nurse & Clinician Notes

- Encounter
- Laboratory
  - Complete Blood Count
  - Hemoglobin A1C
  - Chemistry Panel
  - Full Blood Panel
  - Metabolic Panel
  - Electrolyte Panel
  - Urinalysis

- Pathology
  - Tissue Type
  - Sample Collection
  - Date
  - Stained Samples
  - Genetic Marker(s)
  - Pathology Diagnosis
  - Secondary Findings

- Radiology
  - CT Scans
  - X-ray
  - Ultrasound
  - MRI
  - Radiologist Report
  - Secondary Findings
The power and diversity of EHR studies
Towards a learning health system

A call for deep-learning healthcare

Beau Norgeot, Benjamin S. Glicksberg & Atul J. Butte

Nature Medicine 25, 14–15 (2019) | Download Citation

Fig. 1 | A deep-learning healthcare system. A schematic representation of a deep-learning healthcare system is shown.
Challenges of using EHR data for research

• EHRs are challenging to represent health state
  o heterogeneous
  o noisy
  o incomplete
  o structured / unstructured
  o redundant
  o subject to random errors
  o subject to systematic errors
  o ...and so and so forth
EHR barriers to entry

• Computational

• Domain knowledge:
  • Structure
  • Language

Bodenreider, O (2004): Medical Language System (UMLS): integrating biomedical terminology
Cross-validation & replication in EHR research

Model/Results

EPIC

Cerner

Model/Results
OMOP common data model (CDM)

Language

Structure

Resources:
https://www.ohdsi.org/
http://www.ohdsi.org/web/wiki/doku.php
http://forums.ohdsi.org/
https://github.com/OHDSI/
(most documentation)

Analysis
Scalable and accurate deep learning with electronic health records

Alvin Rajkomar, Eyal Oren, [...] Jeffrey Dean
npj Digital Medicine 1, Article number: 18 (2018)  Download Citation

CDM facilitates cross-validation and reproducibility

FHIREMOP

Association of Hemoglobin A1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin
Analysis From the Observational Health Data Sciences and Informatics Initiative
Rohit Venkatesh, PhD,1,2 Kenneth Jung, PhD,1,2, Alejandro Schuler, MS1, et al.

August 24, 2018

The CDM within the UC system

- Five UC medical centers
- ~14 million unique patients

Network for cross-validation experiments
OMOP CDM across the UC system
The OMOP system is efficient but complicated

- OMOP still requires extensive domain and computational expertise
OHDSI has developed powerful, advanced tools

https://github.com/OHDSI

Open-Source Software

Observational Data Management – tools and processes to standardize the structure and content of healthcare data in preparation for observational analyses, including:

- ATHENA standardized vocabularies
- Common data model and standardized vocabularies specifications
- Extract, transform, and load (ETL) design, development, and testing
- Database profiling and data quality assessment

Clinical Characterization – descriptive analyses to support disease natural history and quality improvement, including:

- Cohort definition and phenotype evaluation
- Patient record profiling
- Study feasibility assessment
- Population summarization and comparison

Population-Level Estimation – epidemiologic designs for estimating average treatment effects for medical product safety surveillance and comparative effectiveness, including:

- Comparative cohort analysis
- Self-controlled case series
- Self-controlled cohort

Patient-level prediction – machine learning methods for precision medicine and disease interception, including:

- Regularized regression
- Random forest
- k-nearest neighbors

https://www.ohdsi.org/analytic-tools/
...that are sometimes *too* advanced for most tasks


ROMOP
a light-weight R package for interfacing with OMOP-formatted Electronic Health Record data

Glicksberg et al. *JAMIA Open* (ooy059)
Goals of ROMOP

1. Automatically connect to OMOP EHR relational database
2. Enable non-technical experts to easily pull data into R-object
3. Facilitate follow-up analyses
What can ROMOP do?

1. Explore CDM fields
2. Generate population statistics
3. Search for patients:
   • Any vocabulary
   • Inclusion/Exclusion criteria
   • Flexible search strategies (e.g., and vs. or)
4. Retrieve all relevant data for patients:
   • Demographics
   • Encounters
   • Clinical
5. Automatically map concepts to ontologies
6. Export search report
Public sandbox server: interactive tutorial

http://romop.ucsf.edu

- 1MM patients from CMS synthesized clinical dataset (DE-SymPUF)

- Package: https://github.com/BenGlicksberg/ROMOP
Data and CDM exploration

Data exploration

- Explore data types in the data ontology

For those unfamiliar with OMOP structure, this function details relevant vocabularies per clinical domain: Condition, Observation, Measurement, Device, Procedure, Drug.

Show data types:

<table>
<thead>
<tr>
<th>Code</th>
<th>Start Over</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><code>showDataTypes()</code></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>domain_id</th>
<th>vocabulary_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>ICD10CM</td>
</tr>
<tr>
<td>Condition</td>
<td>SNOMED</td>
</tr>
<tr>
<td>Condition</td>
<td>ICD9CM</td>
</tr>
<tr>
<td>Device</td>
<td>SNOMED</td>
</tr>
<tr>
<td>Device</td>
<td>HCPCS</td>
</tr>
<tr>
<td>Device</td>
<td>NDC</td>
</tr>
<tr>
<td>Device</td>
<td>SPL</td>
</tr>
<tr>
<td>Drug</td>
<td>NDFRT</td>
</tr>
<tr>
<td>Drug</td>
<td>RxNorm</td>
</tr>
<tr>
<td>Drug</td>
<td>SNOMED</td>
</tr>
</tbody>
</table>

1-10 of 35 rows
Define cohorts/Find patients

Finding cohort/patients

ROMOP has a straightforward yet flexible way to search for patients that takes advantage of the underlying OMOP CDM structure. If the “mapped” option is selected, searching for a broad code like ATC level 3 code A05A (“Bile Therapies”), or even a specific term code like RxNorm code 1544460 for idelalisib, will automatically identify and query for all bottom-level (e.g., idelalisib 150 MG Delayed Release Oral Tablet) codes contained underneath that seed concept. This works by ROMOP first mapping the initial search criteria to a standard concept (SNOMED or RxNorm) and finding all descendants underneath it. This function allows for incorporation of multiple vocabulary types (e.g., ATC and LOINC codes) and codes simultaneously and can support both inclusion and exclusion criteria, if desired. The user can also set the strategy of dealing with criteria, namely either union (i.e., or) or intersection (i.e., and) requirements.

Find all “Type 2 Diabetes Mellitus” patients using ICD10 code (E11):

```r
patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ICD10CM", codes_in = "E11")
```

[1] "5378 patients found that meet the inclusion criteria."

Find all patients prescribed with any “Serotonin receptor antagonists” using ATC code (A03AE):

```r
patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ATC", codes_in = "A03AE")
```

[1] "96 patients found that meet the inclusion criteria."

Find all patients with “Other anxiety disorders” using ICD10 code (F31), but not prescribed with “Clonazepam” using RxNorm code (2598):

```r
patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ICD10CM", codes_in = "F31", strategy_out="mapped", vocabulary_out = "RxNorm", codes_out = "2598")
```

[1] "268 overlapping patients excluded from the original inclusion input based on the exclusion criteria."
[1] "2057 patients found that meet the inclusion criteria."
Extract Data

Retrieved clinical data for pre-defined cohort

Retrieve clinical data for patient ids found from the findPatients function:

Clinical data can also be retrieved for a patient list that is defined using the findPatients function.

```r
patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ATC", codes_in = "A03AE")
ptClinicalData <- getClinicalData(patient_list, declare=FALSE)
head(ptClinicalData$Condition)
```

```
[1] "96 patients found that meet the inclusion criteria."
```

As mentioned, the clinical data are stored as a list of data.tables in the ptClinicalData object.
Summarize cohort demographic information of clinical cohort.

ROMOP provides a function to quickly summarize demographic information for a cohort of interest.
PatientExploreR

dynamic visualization of clinical history in OMOP format

Glicksberg et al. (in revision)
No flexible application exists
Goals

PatientExploreR: dynamic visualization of clinical history

This application allows for flexible searching and extracts patient-level interactive and dynamic reports and visualization of clinical data.

User ID

Password

LOGIN

LOGOUT

Please log-in with your credentials.

Successfully logged in.

First time user? Check out the Help page or start the Tutorial

Patient Finder

Identify a patient to explore: query the EMR for all patients with data a concept or concepts of interest. Can search by Diagnosis, Medication, Procedure, and Lab related concepts. Can further filter patients by demographic features (e.g., age range, self-reported race).

Overall Report

Generate overall report of a selected patient’s clinical history: this report will provide a chronological history of all events of all data modalities (e.g., diseases, medications). Can filter by event type for more focused displays.

Encounter Timeline

Interact and explore a selected patient’s clinical encounter timeline: investigate clinical events by encounter. Selecting an encounter in the timeline will detail all associated clinical events. Can filter by encounter (e.g., Appointment) and visit (e.g., Screening) types.

Data Explorer

Explore patterns of clinical events over time: for a selected patient, can view all data measured for categorical (diseases, medications, procedures) and numeric (labs, vital signs, and flowsheet) types over time. Categorical variables displayed in a timeline and can be filtered for what is shown. Numeric variables are displayed as a timeseries which the user can interact with. Targeted view provides an in-depth graph of one variable at a time while the Multiplex view allows for simultaneous and linked exploration of multiple variables.

Who We Are

Butte Lab, Institute for Computational Health Sciences, UCSF

Contact & Lab Logo/Description
Public Sandbox Server

http://patientexplorer.ucsf.edu

- Synthesized data (no PHI) from CMS
- 1 million patients
- OMOP format
- Open to the public

Code: https://github.com/BenGlicksberg/PatientExploreR
### Criteria (select from table):

<table>
<thead>
<tr>
<th>concept_code</th>
<th>concept_name</th>
<th>domain_id</th>
<th>vocabulary_id</th>
<th>concept_class_id</th>
</tr>
</thead>
<tbody>
<tr>
<td>K51.4</td>
<td>Inflammatory polyp of colon</td>
<td>Condition</td>
<td>ICD10CM</td>
<td>4-char nonbill code</td>
</tr>
<tr>
<td>K51.414</td>
<td>Inflammatory polyp of colon with abscess</td>
<td>Condition</td>
<td>ICD10CM</td>
<td>6-char billing code</td>
</tr>
<tr>
<td>K51.41</td>
<td>Inflammatory polyp of colon with complications</td>
<td>Condition</td>
<td>ICD10CM</td>
<td>5-char nonbill code</td>
</tr>
<tr>
<td>K51.413</td>
<td>Inflammatory polyp of colon with fistula</td>
<td>Condition</td>
<td>ICD10CM</td>
<td>6-char billing code</td>
</tr>
<tr>
<td>K51.412</td>
<td>Inflammatory polyp of colon with intestinal obstruction</td>
<td>Condition</td>
<td>ICD10CM</td>
<td>6-char billing code</td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 64 entries (filtered from 93,463 total entries)

### Selected Criteria:

<table>
<thead>
<tr>
<th>vocabulary</th>
<th>term</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD10CM</td>
<td>K51.4</td>
</tr>
<tr>
<td>ICD10CM</td>
<td>K51.414</td>
</tr>
<tr>
<td>ICD10CM</td>
<td>K51.41</td>
</tr>
<tr>
<td>ICD10CM</td>
<td>K51.413</td>
</tr>
<tr>
<td>ICD10CM</td>
<td>K51.412</td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 64 entries
Automatically generated clinical history

Background:
Status: Alive
Age: 22
Gender: MALE
Ethnicity: Not Hispanic or Latino
Race: Unknown

Clinical Summary:
Earliest encounter: 2017-01-17
Most recent encounter: 2017-07-28
4 encounters: 7
Outpatient encounters: 7
Inpatient encounters: 0

### Observations
- **09-09-17**: Contraceptive use behavior
- **09-09-17**: Drug injection behavior
- **09-09-17**: Hematuria
- **09-09-17**: Creatine kinase [Mass/volume] in Serum or Plasma
- **09-09-17**: C reactive protein [Mass/volume] in Serum or Plasma
- **09-09-17**: Erythrocyte sedimentation rate
- **09-09-17**: Creatinine serum/plasma
- **09-09-17**: Albumin serum/plasma

### Medications
- **09-09-17**: Erythrocyte sedimentation rate
- **09-09-17**: Creatinine serum/plasma
- **09-09-17**: Albumin serum/plasma

# observations: 3
# unique observation concepts: 3
# conditions: 5
# unique condition concepts: 4
# procedures: 0
# unique procedure concepts: 0
# medication prescriptions: 3
# unique medication concepts: 2
# measurements: 40
# unique measurement concepts: 6
# devices: 0
# unique device concepts: 0
### Encounters Timeline: 9000000

#### Plot Encounters:
- None
- Visit Types
- Admitting Concepts
- Discharge Concepts

#### Visit Types
- **OUTPATIENT VISIT**

#### Admitting Concept Type
- NO MATCHING CONCEPT

#### Discharge Concept Type
- NO MATCHING CONCEPT

### Encounter Information:
**Visit Date:** 2017-07-01  
**Visit Type:** Outpatient Visit  
**Visit Admitting Type:** No matching concept  
**Visit Discharge Type:** No matching concept

#### Conditions
<table>
<thead>
<tr>
<th>condition_concept_name</th>
<th>condition_type</th>
<th>condition_status_type</th>
<th>condition_concept_vocabular</th>
<th>condition_concept_code</th>
<th>condition_source_vocabular</th>
<th>condition_source_code</th>
<th>condition_start_date</th>
<th>condition_end_date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergic rhinitis</td>
<td>Primary Condition</td>
<td>SNOEMD</td>
<td>61582004</td>
<td></td>
<td></td>
<td></td>
<td>2017-07-01</td>
<td>2017-07-10</td>
</tr>
</tbody>
</table>

Showing 1 to 1 of 1 entries
Explore Trends in Data/Outcomes (targeted)
Explore Trends in Data/Outcomes (numeric; targeted)
Explore Trends in Data/Outcomes (multiplex)
Explore Trends in Data/Outcomes (multiplex timeline)
How might these tools enable AI-based EHR research?
How are diseases defined using EHR?

Wei et al., JAMIA, 2016
Automated disease cohort selection using word embeddings from Electronic Health Records

Benjamin S. Glicksberg\textsuperscript{1,2*}, Riccardo Miotto\textsuperscript{1,2*}, Kipp W. Johnson\textsuperscript{1,2}, Khader Shameer\textsuperscript{1,2}, Li Li\textsuperscript{1,2}, Rong Chen\textsuperscript{1}, Joel T. Dudley\textsuperscript{1,2}

\textit{Department of Genetics and Genomic Sciences,}\textsuperscript{1} Institute for Next Generation Healthcare\textsuperscript{2}
\textit{Icahn School of Medicine at Mount Sinai}
\textit{Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl.}
New York, NY 10065, USA

* Authors contributed equally
Corresponding author: joel.dudley@mssm.edu

Learning phenotype embeddings

EMR → 

15-day window → shuffle weight 

sentence → word2vec 

Embeddings 

- disease 
- medication 
- lab test 
- procedure 

low-dimensional representation phenotype space

Retrieval and Disease Cohort Identification

C

word2vec

Learning phenotype embeddings
How embeddings organize the phenotype space
How well can we predict...

• Risk for disease
• Disease onset
• Symptom severity
• Treatment response
• Medication adverse events
• Ideal dose of medication
• Symptom flares
• Length of stay in hospital
More representation/data = better reflection of dx

1. **Precision medicine:** finding similar patients to go beyond treating doctor’s, clinic’s, department’s, hospital’s, or even institution’s expertise.

2. **Disease representation in EHR:** electronic phenotyping algorithms might not be fully generalizable. Building as a “meta” signature will be more robust.

3. **Prediction:** training and testing models across multiple institutions, alone and in conjunction, will enable identifying ideal strategies.

4. **Multi-omic factors:** incorporating genetics and environmental data (e.g., pollution) can help pinpoint etiology and discern GxE interactions.
Acknowledgements

Boris Oskotsky, PhD
Vivek Rudrapatna, MD, PhD
Debajyoti Datta, MD, PhD
Beau Norgeot
Nadav Rappoport, PhD
Ted Goldstein, PhD
Atul Butte, MD, PhD

Eugenia Rutenberg, MBA
Angelo Pelonero
Angela Rizk-Jackson, PhD
Sharat Israni, PhD

Nicholas Giangreco
Phyllis Thangaraj
Nicholas Tatonetti, PhD

Li Li, MD
Khader Shameer, PhD
Riccardo Miotto, PhD
Kipp Johnson
Marcus Badgeley
Mark Shervey
Rong Chen, PhD
Joel Dudley, PhD

Dana Ludwig, MD
Remi Frazier
Nelson Lee
Rick Larsen

Community and Developers