Apps To Enable AI In EHR Research

Ben Glicksberg, PhD
Butte Lab
Bakar Computational Health Sciences Institute
University of California, San Francisco (UCSF)

@BenGlicksberg
Clinical Informatics in the era of big data

"The Quantified Self"

- Input predisposition
- Environment
- Family History
- Genetic Architecture
- Demographics
  - Sex
  - Race
  - Age
- Gene x Environment Interaction
- Exposures
- Lifestyle
- Genes
- Toxins
- Chemicals
- Therapeutics
- Patient Monitoring
- System
- Artificial Intelligence
- Machine Learning
- Output
- Interventional Therapies
- Risk Factor Identification
- Disease Monitoring
- Patient Monitoring
- Therapeutic Stratification

Electronic Health Records

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

Demographics
- 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
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
OMOP common data model (CDM)

Analysis

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)
CDM facilitates cross-validation and reproducibility

Scalable and accurate deep learning with electronic health records

All available data for each patient is converted to events recorded in containers based on the Fast Healthcare Interoperability Resource (FHIR) specification.

The FHIR resources are placed in temporal order, depicting all events recorded in the EHR (i.e., timeline). The deep learning model uses this full history to make each prediction.

OMOP

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

August 24, 2018

Rohit Vashist, PhD; Kenneth Jung, PhD; Alejandro Schuler, MS; et al

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

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
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 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>showDataTypes()</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>&lt;chr&gt;</td>
<td>&lt;chr&gt;</td>
</tr>
<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 idealisib, will automatically identify and query for all bottom-level (e.g., idealisib 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.

1. Find all "Type 2 Diabetes Mellitus" patients using ICD10 code (E11):

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

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

2. Find all patients prescribed with any "Serotonin receptor antagonists" using ATC code (A03AE):

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

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

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

```
patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ICD10CM", codes_in = "F31", strategy_out="mapped", codes_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

- Retrieve 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."
```

<table>
<thead>
<tr>
<th>condition_concept_vocabulary</th>
<th>condition_concept_code</th>
<th>condition_concept_name</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNOMED</td>
<td>40257000</td>
<td>Contusion of shoulder region</td>
</tr>
<tr>
<td>SNOMED</td>
<td>40257000</td>
<td>Contusion of shoulder region</td>
</tr>
<tr>
<td>SNOMED</td>
<td>35678005</td>
<td>Multiple joint pain</td>
</tr>
<tr>
<td>SNOMED</td>
<td>44465007</td>
<td>Sprain of ankle</td>
</tr>
<tr>
<td>SNOMED</td>
<td>95210003</td>
<td>Plasma cell leukemia</td>
</tr>
<tr>
<td>SNOMED</td>
<td>11437003</td>
<td>Contusion of back</td>
</tr>
</tbody>
</table>

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

Summarize demographic information of clinical cohort

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

Summarize demographic information for patient ids found from the findPatients function:

```r
patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ATC", codes_in = "AB34E")
ptDemo <- getDemographics(patient_list, declare=FALSE)
summarizeDemographics(ptDemo)
```

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

# of patients: 96
Mean age: 79.375
Median age: 82.5
STD age: 14.145
Status breakdown:

Status n proportion
1: Alive 94 0.9791667
2: Deceased 2 0.0208333

Gender breakdown:

Gender n proportion
1: FEMALE 61 0.6354167
2: MALE 35 0.3645833

Race breakdown:

Race n proportion
1: Black or African American 7 0.0739167
2: Unknown 9 0.0937500
3: White 80 0.8333333

Ethnicity breakdown:

Ethnicity n proportion
1: Hispanic or Latino 8 0.0833333
2: Not Hispanic or Latino 91 0.9166667
PatientExploreR

dynamic visualization of clinical history in OMOP format

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

**EPIC**

**OMOP**

**PatientExploreR**: dynamic visualization of clinical history

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

- **User ID**: glicksberg
- **Password**: ********

Please log-in with your credentials.

Successfully logged in.

- **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., disease, 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
Patient Finder

Search for patients directly or based on clinical criteria (e.g., Condition ICD-10CM code). By selecting 'Criteria', all available ontologies will be displayed per modality which the user can use for searching. This will load demographic information for matching patients to allow for further refining.

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 92,463 total entries)
Automatically generated clinical history

### Background:
- **Status:** Alive
- **Age:** 22
- **Age of Death:** NA
- **Ethnicity:** Not Hispanic or Latino
- **Race:** Unknown
- **Gender:** MALE

### Clinical Summary:
- **Earliest encounter:** 2017-01-17
- **Most recent encounter:** 2017-01-28

#### Data Modalities
- **Observations:** 3 items selected
- **Conditions:** 4 items selected
- **Procedures:** 0 items selected
- **Devices:** 0 items selected

#### Observations
- **2017-01-17**: Contraceptive use behavior
- **2017-01-17**: Drug injection behavior
- **2017-01-17**: Hemocult
- **2017-01-17**: Calprotectin (Mass/mass) in Stool
- **2017-01-17**: C reactive protein (Mass/volume) in Serum or Plasma
- **2017-01-17**: Erythrocyte sedimentation rate
- **2017-01-17**: Creatinine serum/plasma
- **2017-01-17**: Albumin serum/plasma
- **2017-04-15**: Keloid scar
- **2017-07-01**: Tobacco use and exposure

Showing 1 to 10 of 51 entries
Data Explorer: 9000000

Explore all clinical events over the patient's history. The user can explore both categorical (Conditions, Medications, Procedures, or Devices) or numeric (Measurement or Observation) data. For categorical data, the events are visualized in an interactive timeline and the user can select which events to show. Further, diseases may be explored at different levels (Disease Name, ICD 9 or 10). For numeric data types, the events (e.g., WBC for Labs) are displayed as a table with # of measurements recorded. The user can select an event of interest which will display as an interactive timeseries plot.

Visit Occurrence ID for Condition: 9000002
Condition Window: 2017-07-06 to 2017-07-21
Condition Status Type: NA
Condition Standardized Name Selected: Ulcerative colitis
Condition Standardized Vocabulary: SNOMED
Condition Standardized Vocabulary Code: 64766004
Condition Source Value: NA
Condition Source Vocabulary: NA
Condition Source Vocabulary Code: NA
Explore Trends in Data/Outcomes (numeric; targeted)

<table>
<thead>
<tr>
<th>Measurement Concept Name</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albumin serum/plasma</td>
<td>7</td>
</tr>
<tr>
<td>Calprotectin [Mass/mass] in Stool</td>
<td>7</td>
</tr>
<tr>
<td>C reactive protein [Mass/volume] in Serum or Plasma</td>
<td>7</td>
</tr>
<tr>
<td>Erythrocyte sedimentation rate</td>
<td>7</td>
</tr>
<tr>
<td>Hematocrit</td>
<td>7</td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 6 entries
Explore Trends in Data/Outcomes (multiplex)
Explore Trends in Data/Outcomes (multiplex timeline)
How might these tools enable AI-based EHR research?
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