Apps To Enable AI In EHR Research

Butte Lab

UCSF Bakar Computational Health Sciences Institute

Ben Glicksberg, PhD

Butte Lab

Bakar Computational Health Sciences Institute

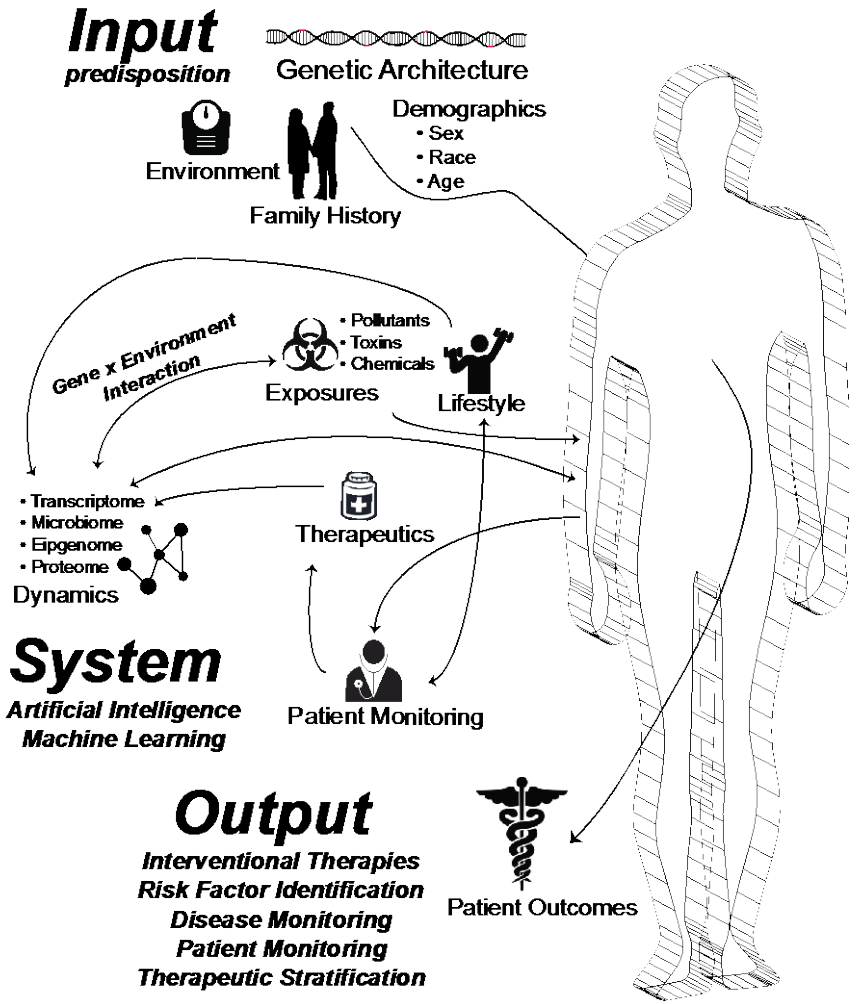
University of California, San Francisco (UCSF)

UCSF University of California San Francisco

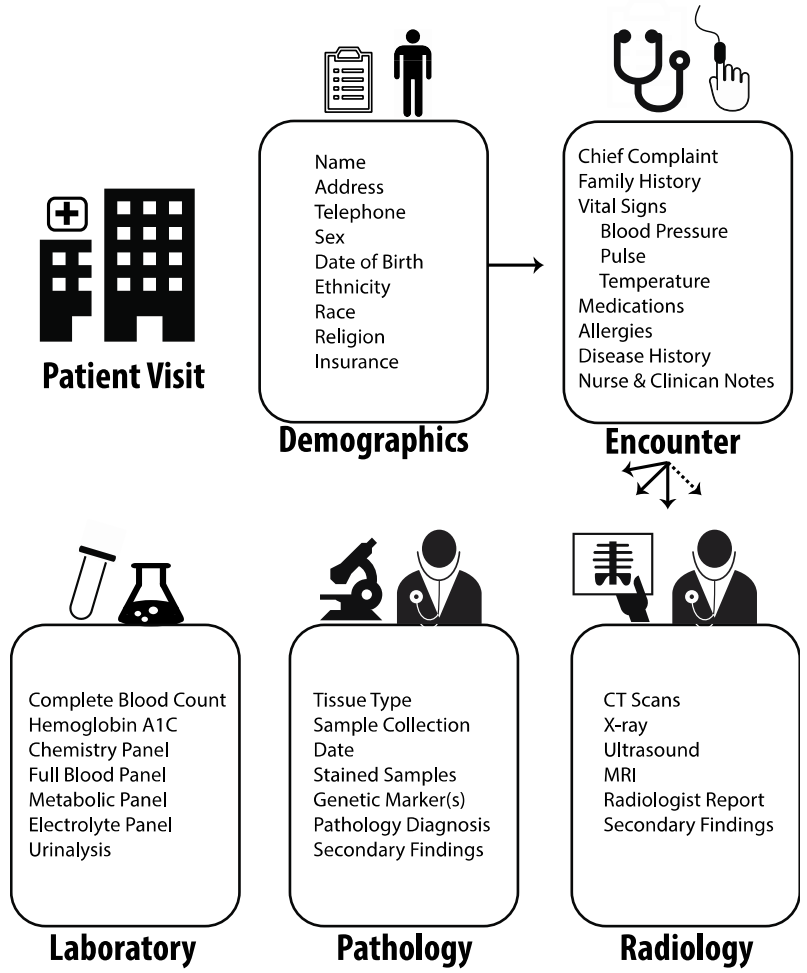
@BenGlicksberg

Clinical Informatics in the era of big data

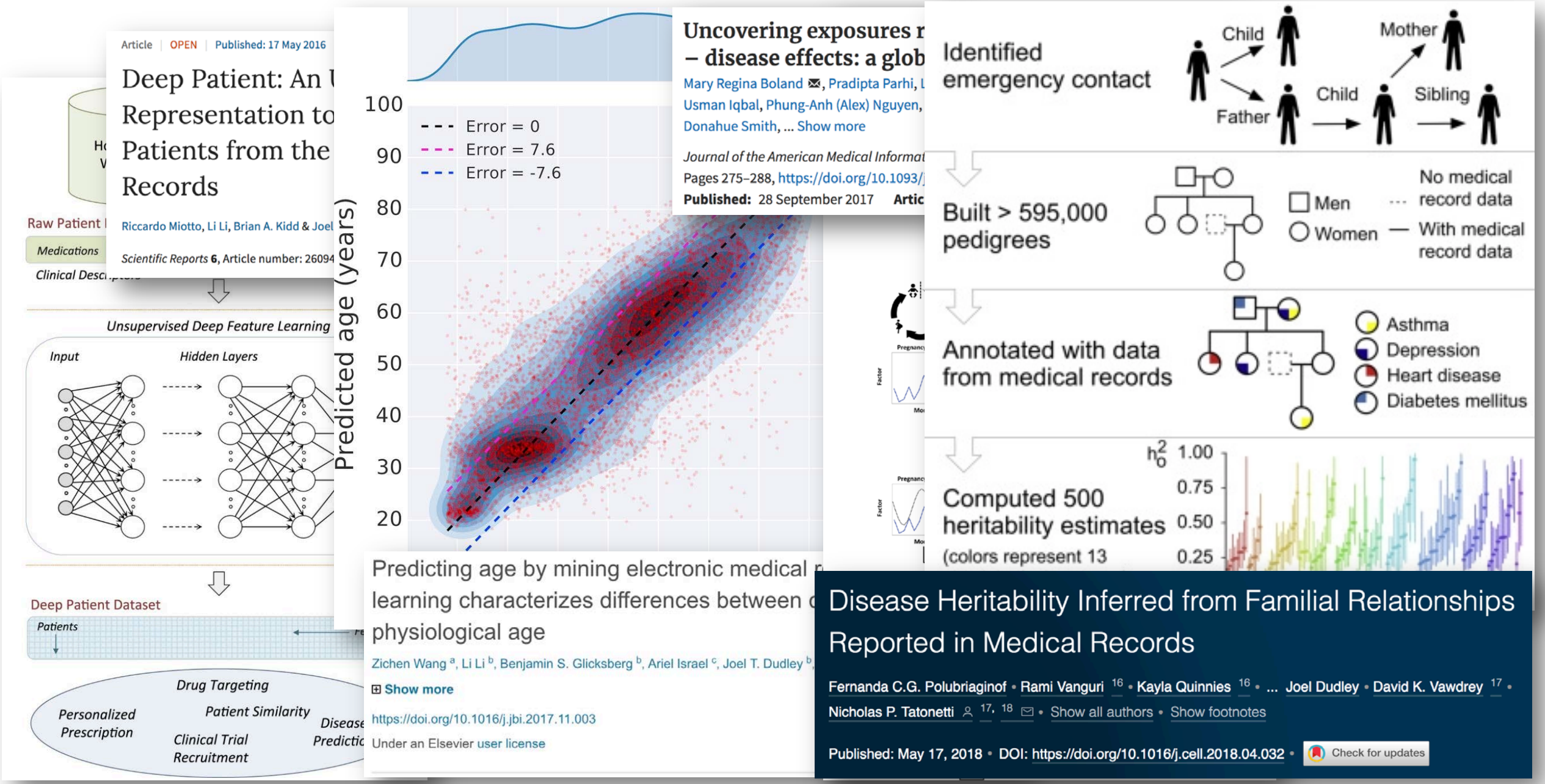
"The Quantified Self"



Electronic Health Records



The power and diversity of EHR studies



Towards a learning health system

nature medicine

Comment | Published: 07 January 2019

A call for deep-learning healthcare

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

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

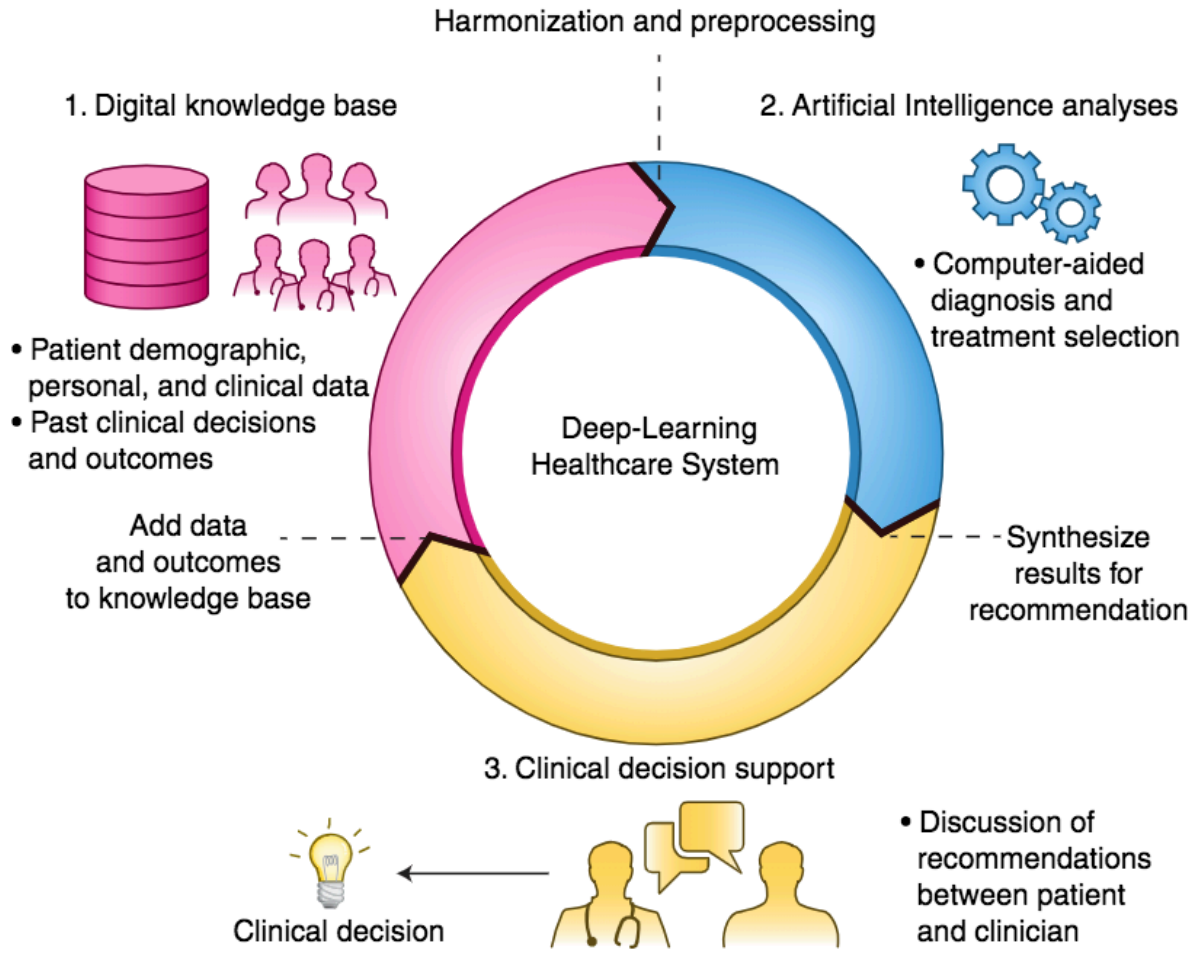


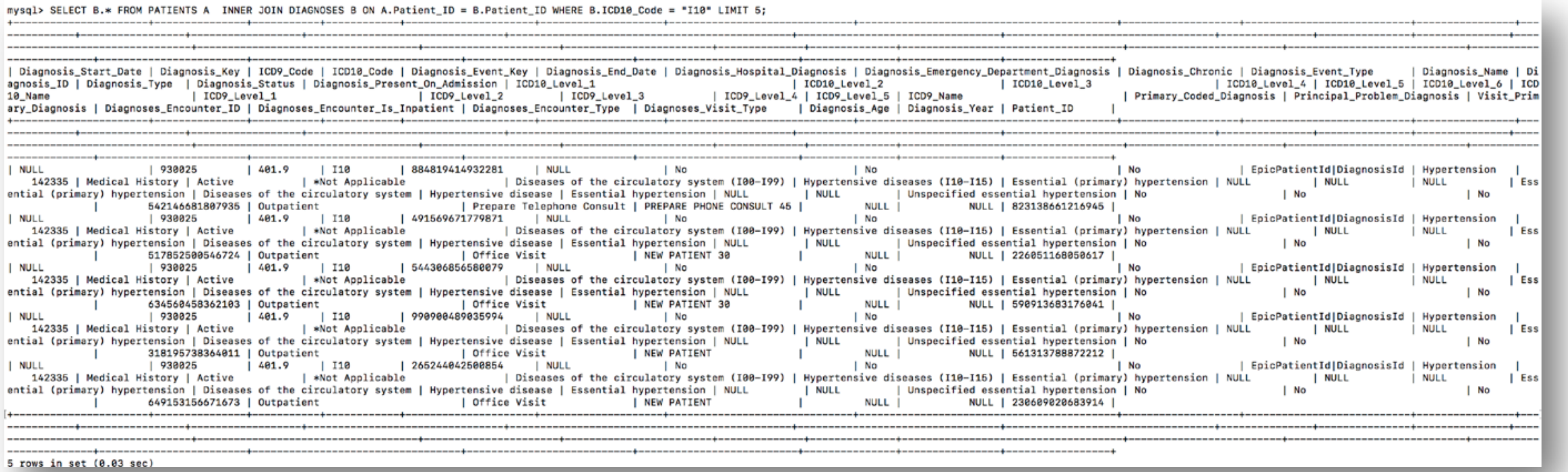
Fig. 1 | A deep-learning healthcare system. A schematic representation of a deep-learning healthcare system is shown.

Challenges of using EHR data for research

- EHRs are challenging to represent health state
  - heterogeneous
  - noisy
  - incomplete
  - structured / unstructured
  - redundant
  - subject to random errors
  - subject to systematic errors
  - *…and so and so forth*

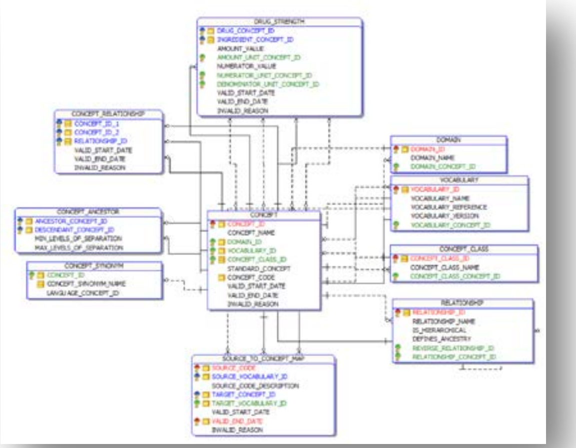
EHR barriers to entry

- Computational

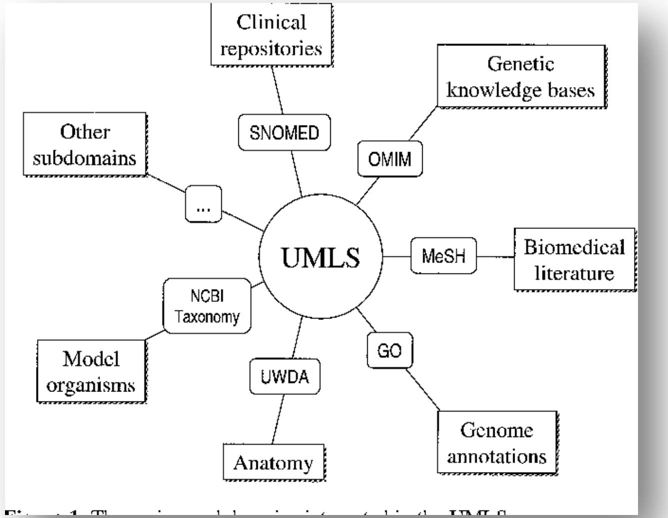


```
mysql> SELECT B.* FROM PATIENTS A  INNER JOIN DIAGNOSES B ON A.Patient_ID = B.Patient_ID WHERE B.ICD10_Code = "I10" LIMIT 5;
| Diagnosis_Start_Date | Diagnosis_Key | ICD9_Code | ICD10_Code | Diagnosis_Event_Key | Diagnosis_End_Date | Diagnosis_Hospital_Diagnosis | Diagnosis_Emergency_Department_Diagnosis | Diagnosis_Chronic | Diagnosis_Event_Type | Diagnosis_Name | Di
agnosis_ID | Diagnosis_Type | Diagnosis_Status | Diagnosis_Present_On_Admission | ICD10_Level_1 | ICD10_Level_2 | ICD10_Level_3 | ICD10_Level_4 | ICD10_Level_5 | ICD10_Level_6 | ICD
10_Name | ICD9_Level_1 | ICD9_Level_2 | ICD9_Level_3 | ICD9_Level_4 | ICD9_Level_5 | ICD9_Name | Primary_Coded_Diagnosis | Principal_Problem_Diagnosis | Visit_Prim
ary_Diagnosis | Diagnoses_Encounter_ID | Diagnoses_Encounter_Is_Inpatient | Diagnoses_Encounter_Type | Diagnoses_Visit_Type | Diagnosis_Age | Diagnosis_Year | Patient_ID |
| NULL | 930025 | 401.9 | I10 | 884819414932281 | NULL | No | No | EpicPatientId|DiagnosisId | Hypertension |
  142335 | Medical History | Active | *Not Applicable | Diseases of the circulatory system (I00-I99) | Hypertensive diseases (I10-I15) | Essential (primary) hypertension | NULL | NULL | NULL | Ess
ential (primary) hypertension | Diseases of the circulatory system | Hypertensive disease | Essential hypertension | NULL | NULL | NULL | Unspecified essential hypertension | No | No | No
  | 542146681807935 | Outpatient | Prepare Telephone Consult | PREPARE PHONE CONSULT 45 | NULL | NULL | 823138661216945 |
| NULL | 930025 | 401.9 | I10 | 491569671779871 | NULL | No | No | EpicPatientId|DiagnosisId | Hypertension |
  142335 | Medical History | Active | *Not Applicable | Diseases of the circulatory system (I00-I99) | Hypertensive diseases (I10-I15) | Essential (primary) hypertension | NULL | NULL | NULL | Ess
ential (primary) hypertension | Diseases of the circulatory system | Hypertensive disease | Essential hypertension | NULL | NULL | NULL | Unspecified essential hypertension | No | No | No
  | 517852500546724 | Outpatient | Office Visit | NEW PATIENT 30 | NULL | NULL | 226051168050617 |
| NULL | 930025 | 401.9 | I10 | 544306856580079 | NULL | No | No | EpicPatientId|DiagnosisId | Hypertension |
  142335 | Medical History | Active | *Not Applicable | Diseases of the circulatory system (I00-I99) | Hypertensive diseases (I10-I15) | Essential (primary) hypertension | NULL | NULL | NULL | Ess
ential (primary) hypertension | Diseases of the circulatory system | Hypertensive disease | Essential hypertension | NULL | NULL | NULL | Unspecified essential hypertension | No | No | No
  | 634560458362103 | Outpatient | Office Visit | NEW PATIENT 30 | NULL | NULL | 590913683176041 |
| NULL | 930025 | 401.9 | I10 | 990900489035994 | NULL | No | No | EpicPatientId|DiagnosisId | Hypertension |
  142335 | Medical History | Active | *Not Applicable | Diseases of the circulatory system (I00-I99) | Hypertensive diseases (I10-I15) | Essential (primary) hypertension | NULL | NULL | NULL | Ess
ential (primary) hypertension | Diseases of the circulatory system | Hypertensive disease | Essential hypertension | NULL | NULL | NULL | Unspecified essential hypertension | No | No | No
  | 318195738364011 | Outpatient | Office Visit | NEW PATIENT | NULL | NULL | 561313788872212 |
| NULL | 930025 | 401.9 | I10 | 265244042500854 | NULL | No | No | EpicPatientId|DiagnosisId | Hypertension |
  142335 | Medical History | Active | *Not Applicable | Diseases of the circulatory system (I00-I99) | Hypertensive diseases (I10-I15) | Essential (primary) hypertension | NULL | NULL | NULL | Ess
ential (primary) hypertension | Diseases of the circulatory system | Hypertensive disease | Essential hypertension | NULL | NULL | NULL | Unspecified essential hypertension | No | No | No
  | 649153156671673 | Outpatient | Office Visit | NEW PATIENT | NULL | NULL | 230609020683914 |
5 rows in set (0.03 sec)
```

- Domain knowledge:
  - Structure

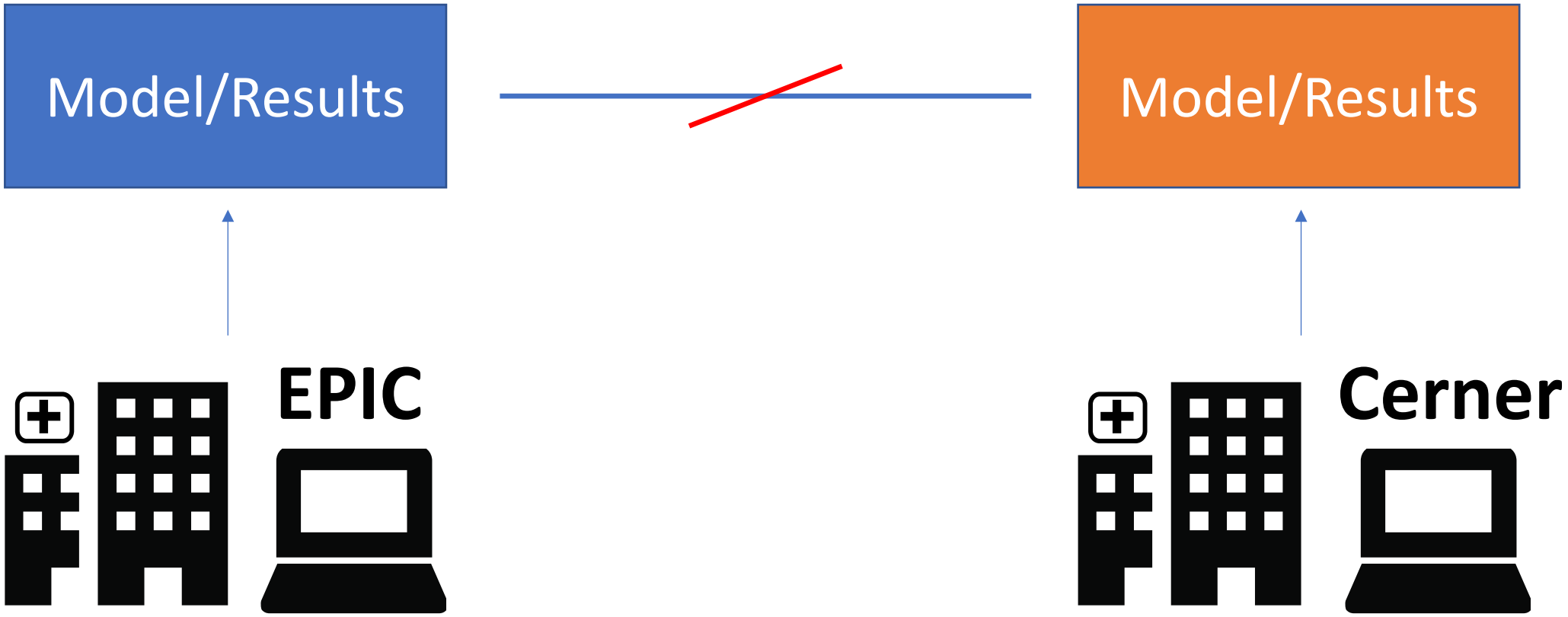


- Language



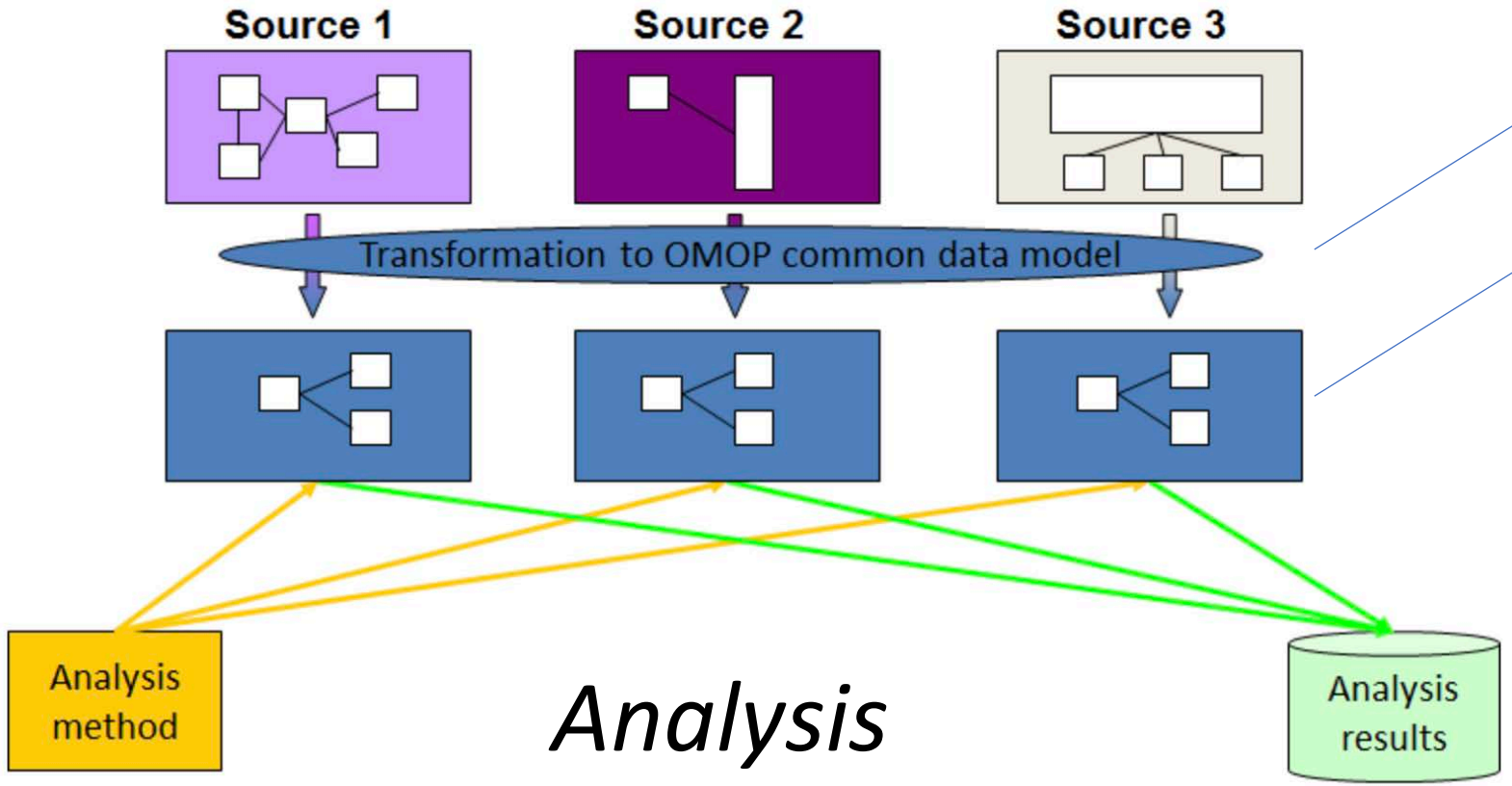
Bodenreider, O (2004): Medical Language System (UMLS) : integrating biomedical terminology

Cross-validation & replication in EHR research



OMOP common data model (CDM)

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS



Language

Structure

Resources:

https://www.ohdsi.org/
http://www.ohdsi.org/web/wiki/doku.php
http://forums.ohdsi.org/

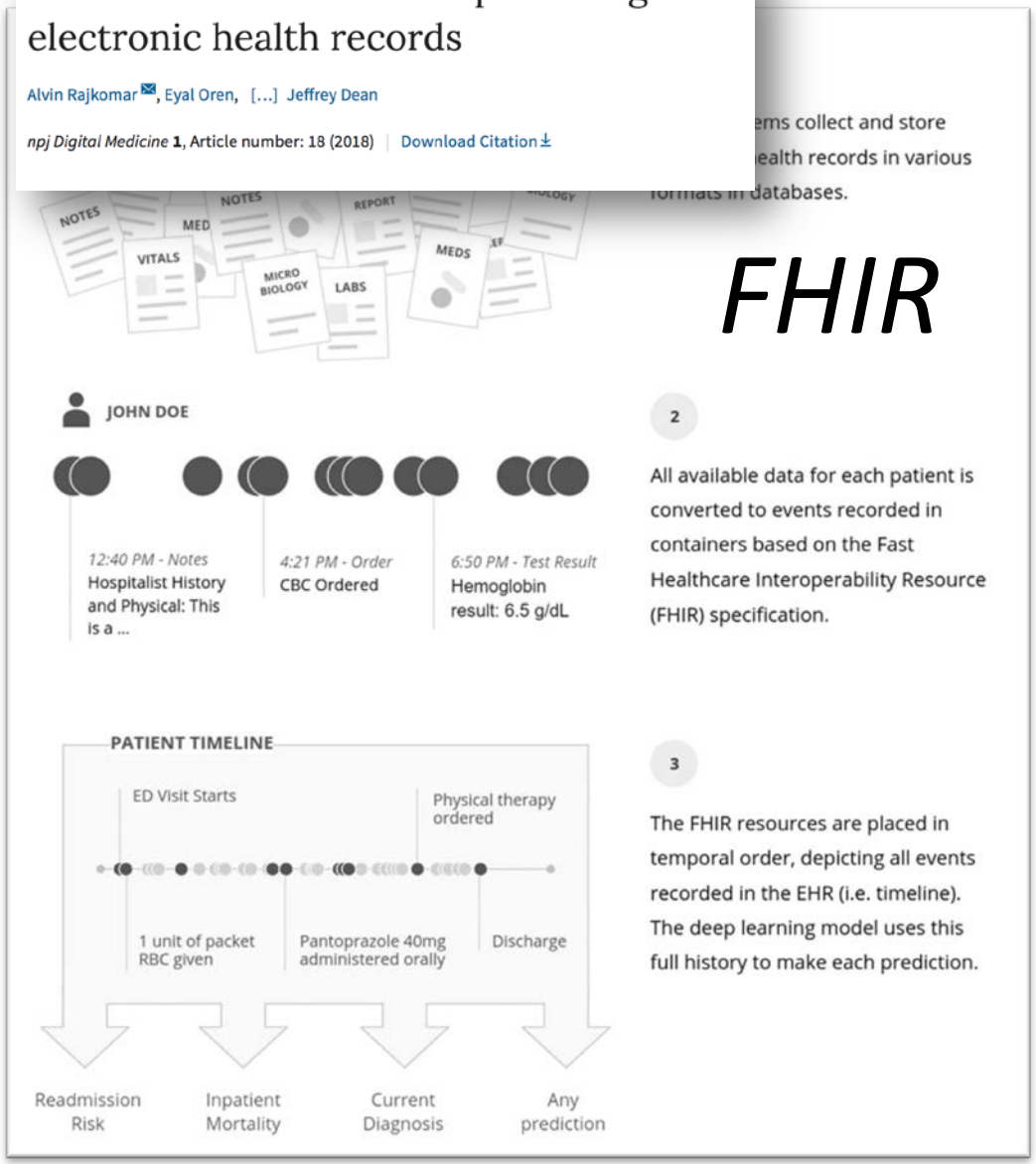
https://github.com/OHDSI/
(most documentation)

CDM facilitates cross-validation and reproducibility

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar, Eyal Oren, [...] Jeffrey Dean

npj Digital Medicine **1**, Article number: 18 (2018) | Download Citation



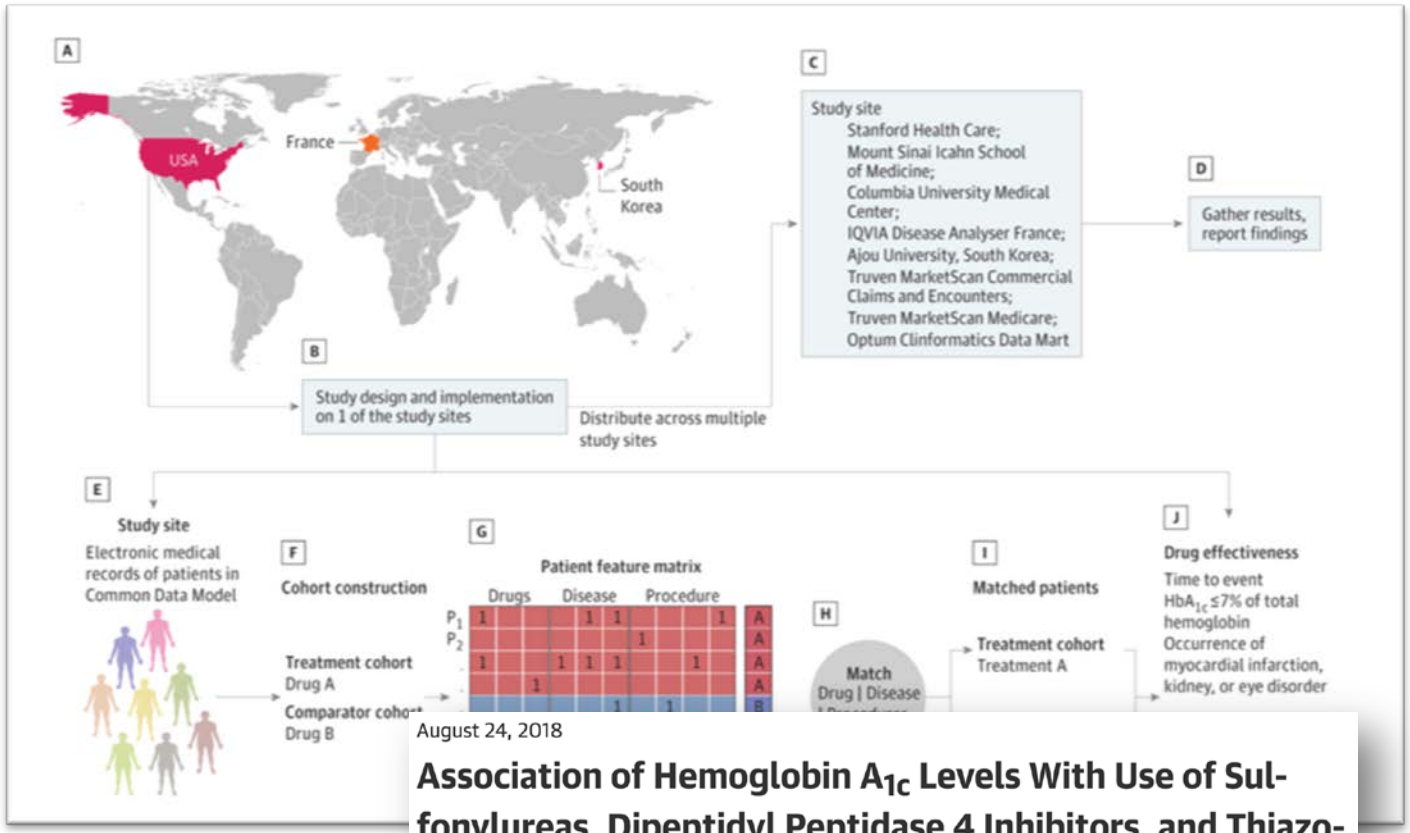
FHIR

2

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

3

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

August 24, 2018

Association of Hemoglobin A_{1c} Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin

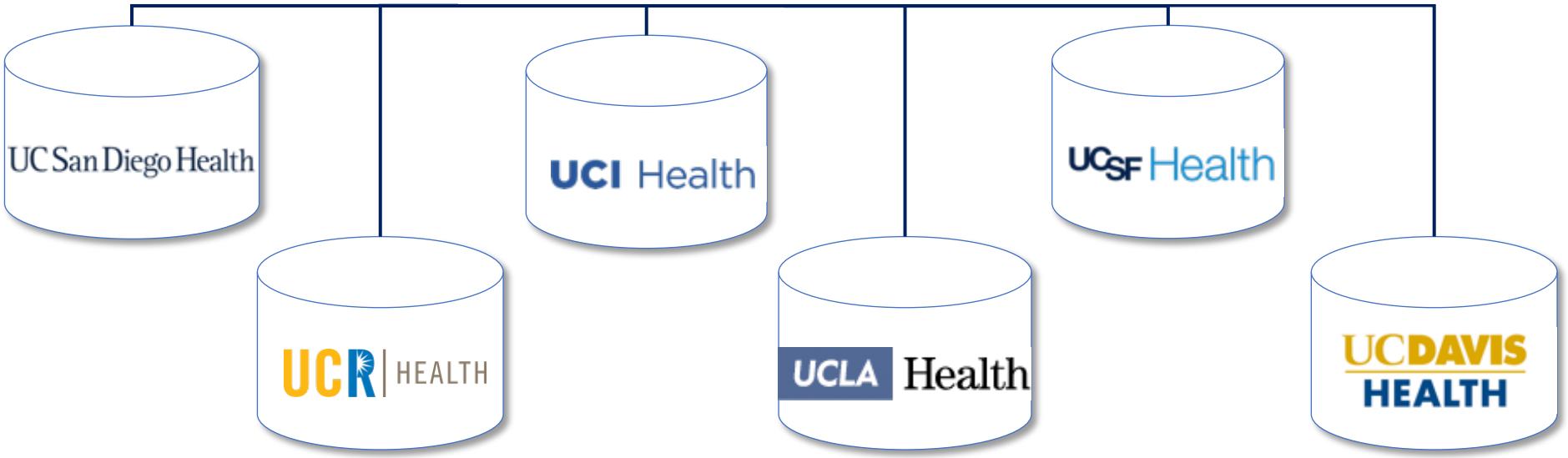
Analysis From the Observational Health Data Sciences and Informatics Initiative

Rohit Vashisht, PhD[1,2]; Kenneth Jung, PhD[1,2]; Alejandro Schuler, MS[1,2]; et al

» Author Affiliations | Article Information

JAMA Netw Open. 2018;1(4):e181755. doi:10.1001/jamanetworkopen.2018.1755

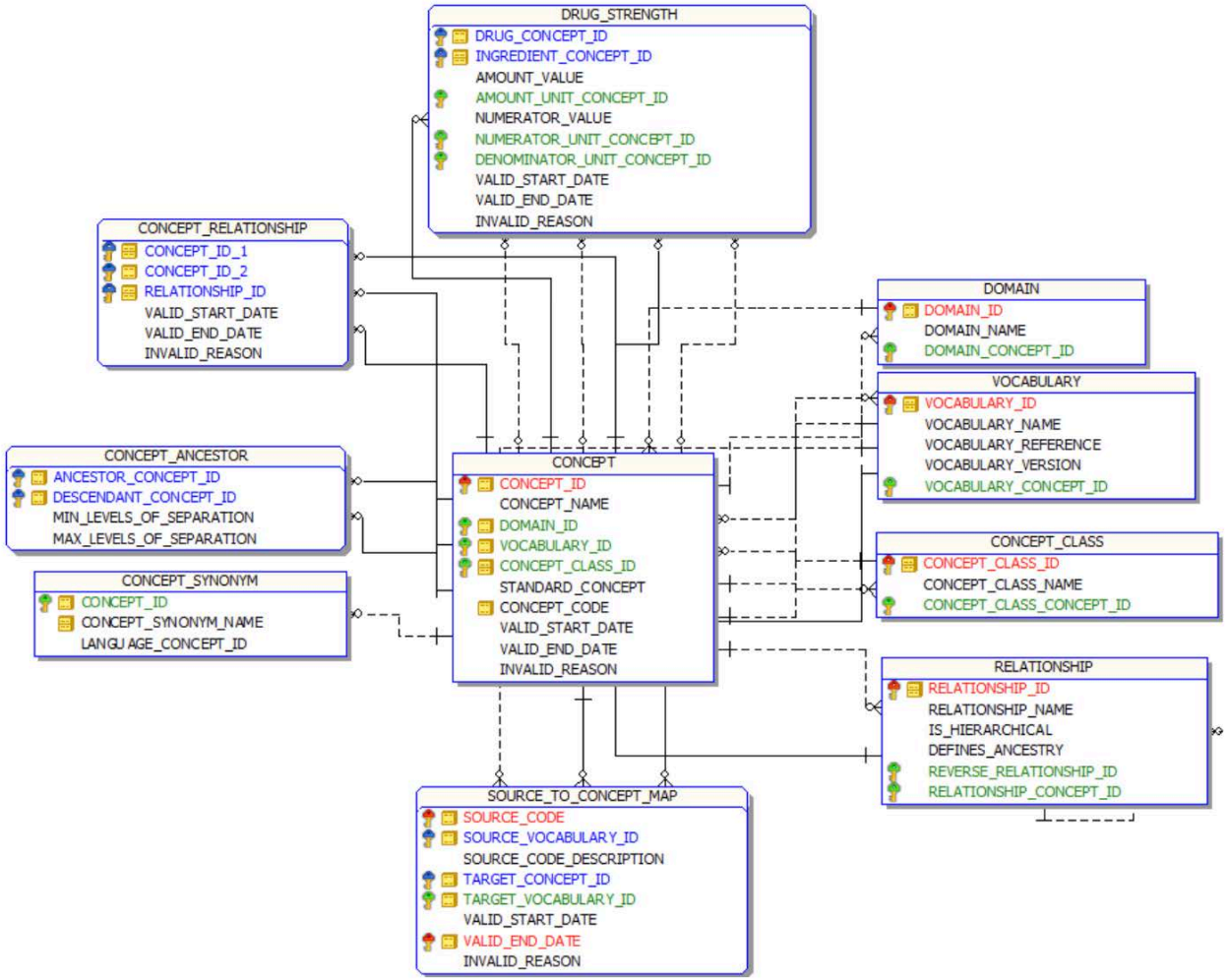
OMOP CDM across the UC system

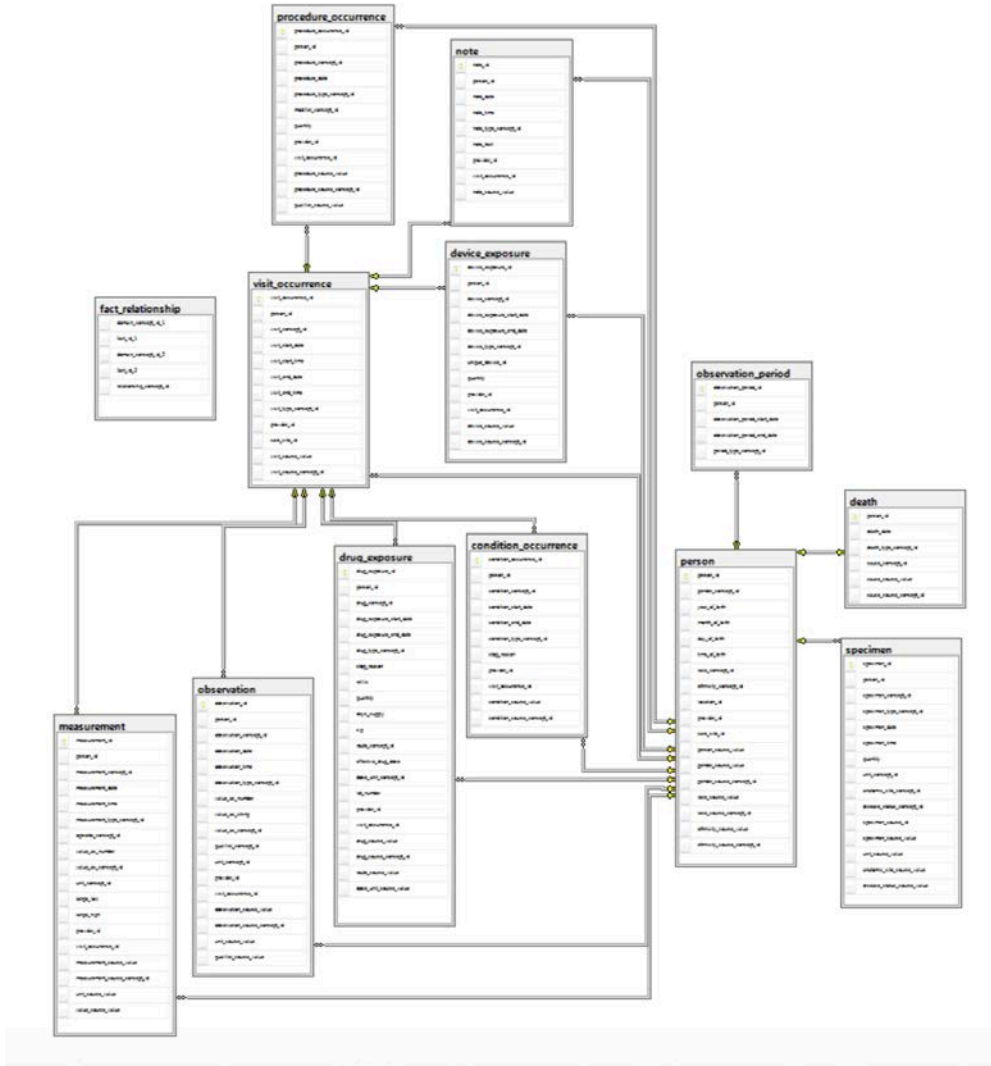


UC Health Data Analytics Platform

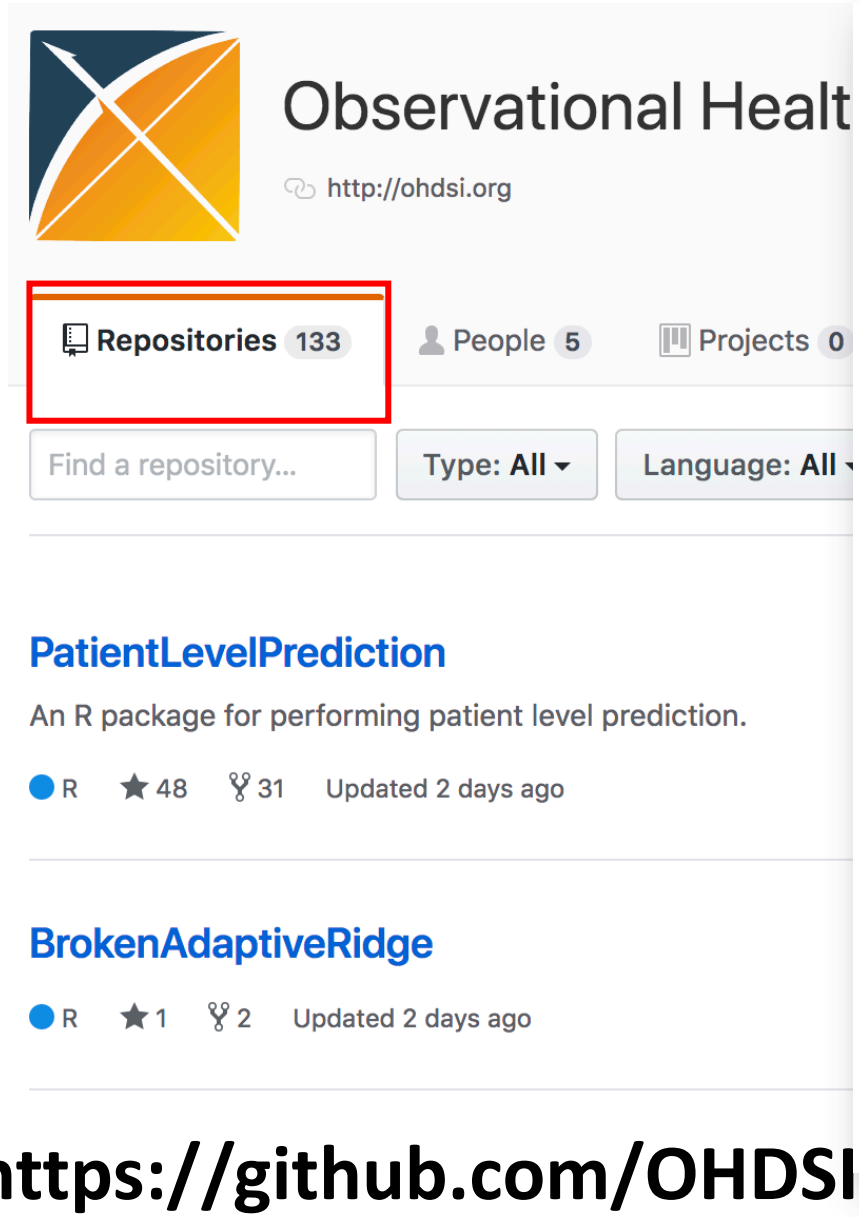
The OMOP system is efficient but complicated

- 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:

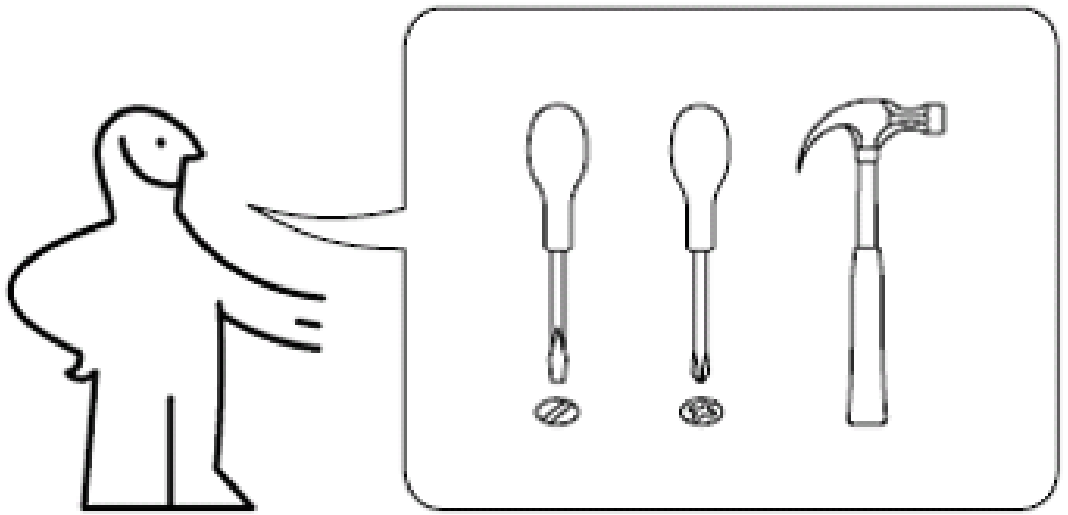
- Comparative cohort analysis
- Self-controlled case series
- Self-controlled cohort

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



http://remembar.me/wp-content/uploads/2018/07/garage-pegboard-organization-interior-furniture-full-image-for-tool-storage-special-tools-and-ideas.jpg

https://www.ikea.com/ms/en_CA/customer_service/assembly_instructions/assembly_instructions1.html

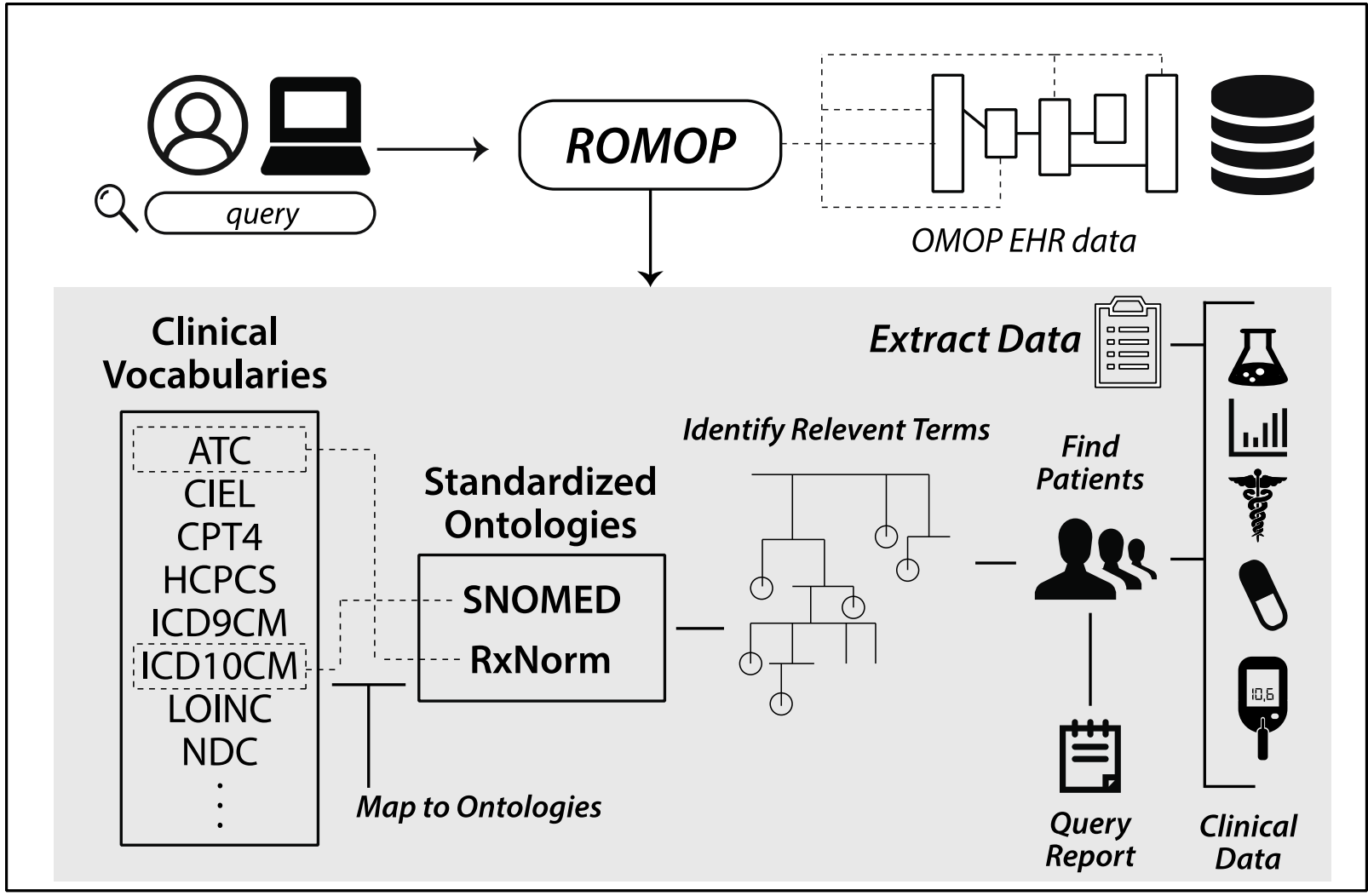
ROMOP

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

Glicksberg et al. *JAMIA Open* (ooy059)

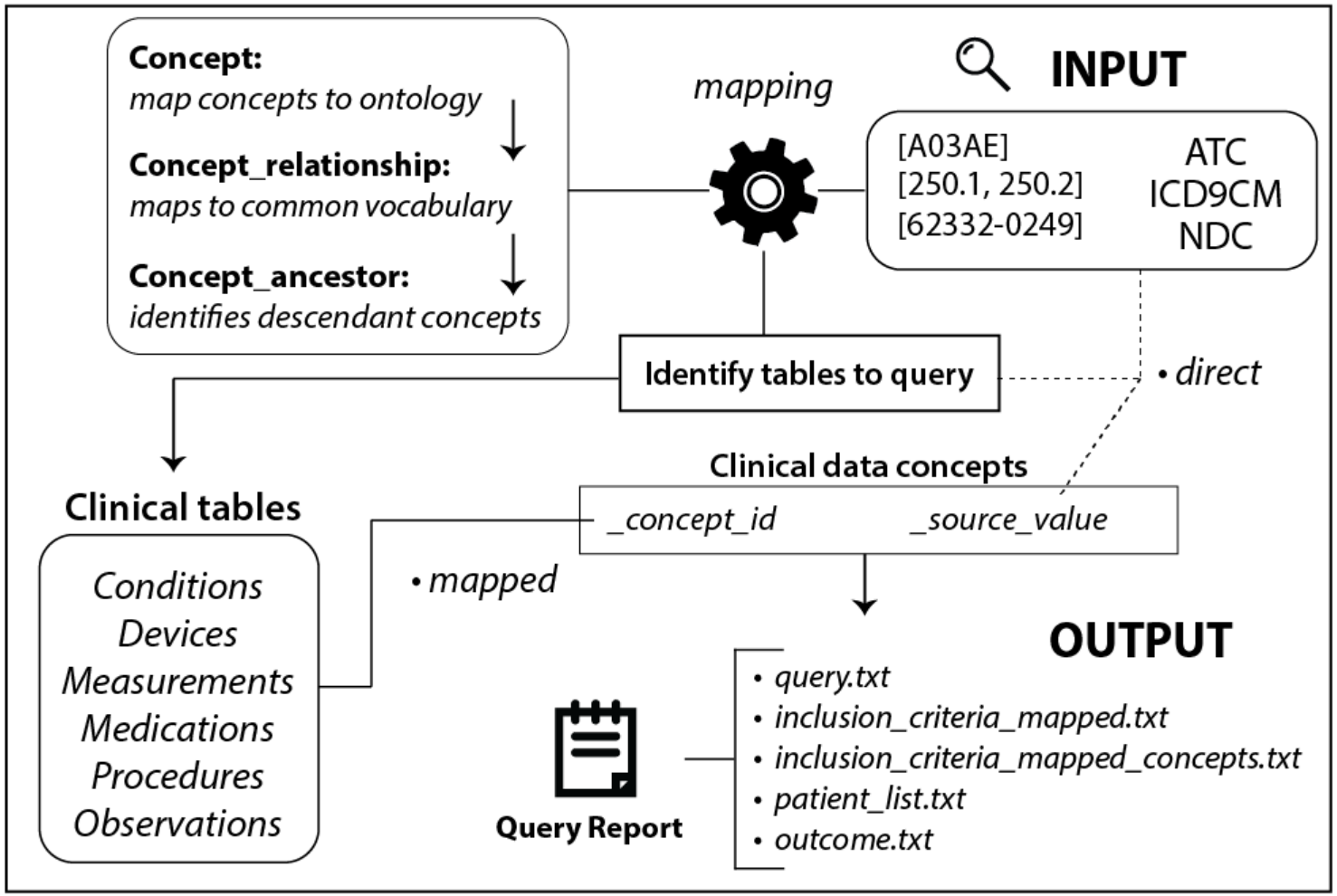
Goals of ROMOP

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



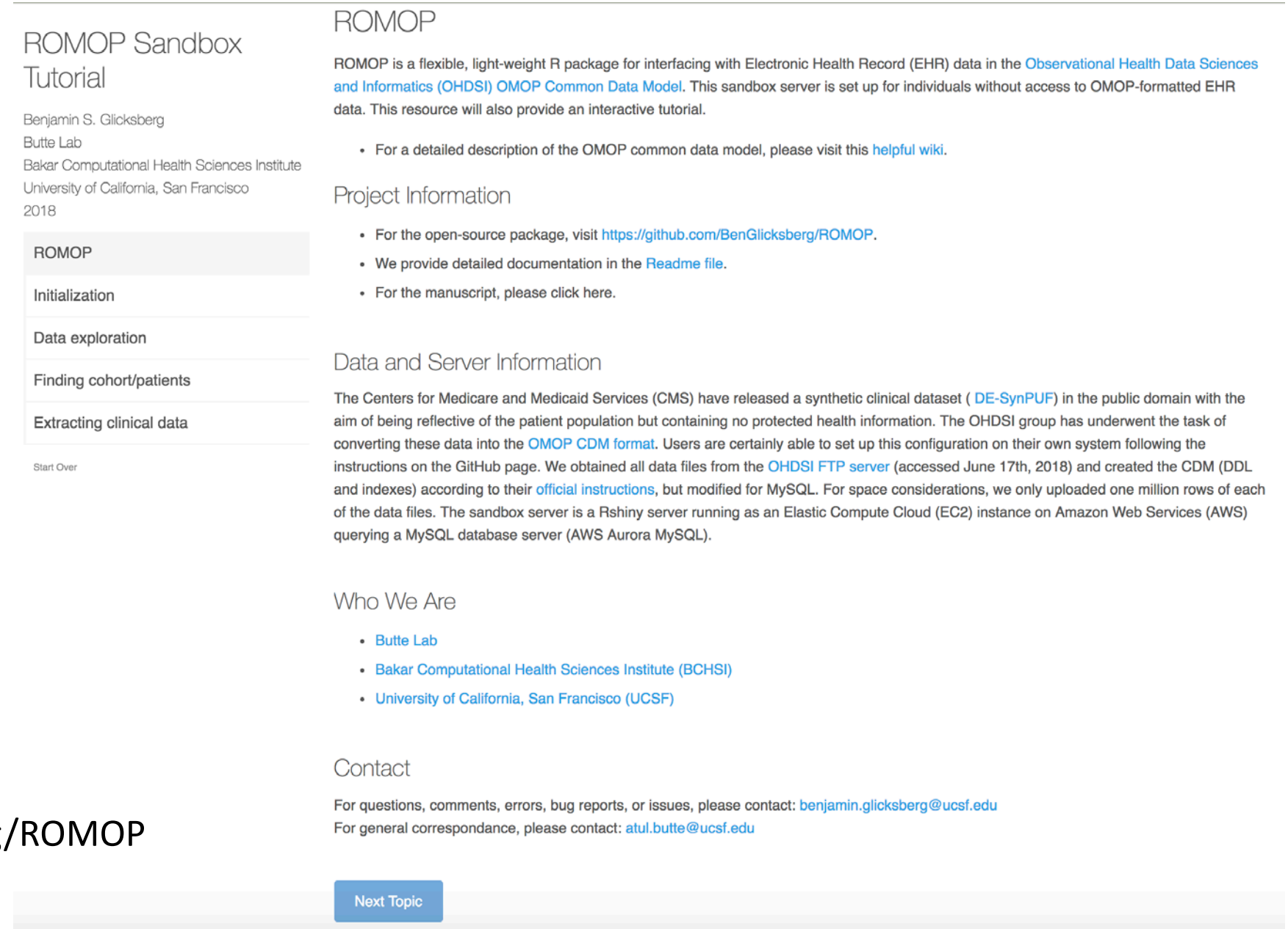
Public sandbox server: interactive tutorial

http://romop.ucsf.edu

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

- Package:

https://github.com/BenGlicksberg/ROMOP



Data and CDM exploration

ROMOP Sandbox Tutorial

Benjamin S. Glicksberg
Butte Lab
Bakar Computational Health Sciences Institute
University of California, San Francisco

2018

- ROMOP
- Initialization
- Data exploration
- Finding cohort/patients
- Extracting clinical data

Start Over

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:

Code | Start Over | Run Code

```
1 showDataTypes()
2
3
```

domain_id <chr>	vocabulary_id <chr>
Condition	ICD10CM
Condition	SNOMED
Condition	ICD9CM
Device	SNOMED
Device	HCPCS
Device	NDC
Device	SPL
Drug	NDFRT
Drug	RxNorm
Drug	SNOMED

1-10 of 35 rows | Previous 1 2 3 4 Next

Define cohorts/Find patients

ROMOP Sandbox Tutorial

Benjamin S. Glicksberg
Butte Lab
Bakar Computational Health Sciences Institute
University of California, San Francisco

2018

Start Over

Finding cohort/patients

ROMOP has a straight-forward yet flexible ways 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):

Code | Start Over | Run Code

```
1 patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ICD10CM", codes_in = "E11")
2
3
```

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

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

Code | Start Over | Run Code

```
1 patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ATC", codes_in = "A03AE")
2
3
```

```
[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):

Code | Start Over | Run Code

```
1 patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ICD10CM", codes_in = "F41", strategy_out="mapped", \
2
3
```

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

Previous Topic | Next Topic

Extract Data

ROMOP Sandbox Tutorial

Benjamin S. Glicksberg
Butte Lab
Bakar Computational Health Sciences Institute
University of California, San Francisco

2018

Start Over

✓ 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.

Code | Start Over | Run Code

```
1 patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ATC", codes_in = "A03AE")
2
3 ptClinicalData <- getClinicalData(patient_list, declare=FALSE)
4
5 head(ptClinicalData$Condition)
```

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

condition_concept_vocabulary <chr>	condition_concept_code <chr>	condition_concept_name <chr>
SNOMED	40257000	Contusion of shoulder region
SNOMED	40257000	Contusion of shoulder region
SNOMED	35678005	Multiple joint pain
SNOMED	44465007	Sprain of ankle
SNOMED	95210003	Plasma cell leukemia
SNOMED	11437003	Contusion of back

6 rows | 5-7 of 12 columns

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

Summarize cohort

ROMOP Sandbox Tutorial

Benjamin S. Glicksberg
Butte Lab
Bakar Computational Health Sciences Institute
University of California, San Francisco

2018

Start Over

✓ 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:

Code | Start Over | Run Code

```
1 patient_list <- findPatients(strategy_in="mapped", vocabulary_in = "ATC", codes_in = "A03AE")
2
3 ptDemo <- getDemographics(patient_list, declare=FALSE)
4
5 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 0.97916667
2: Deceased  2 0.02083333
```

```
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 0.07291667
2:                  Unknown  9 0.09375000
3:                    White 80 0.83333333
```

```
Ethnicity breakdown:
```

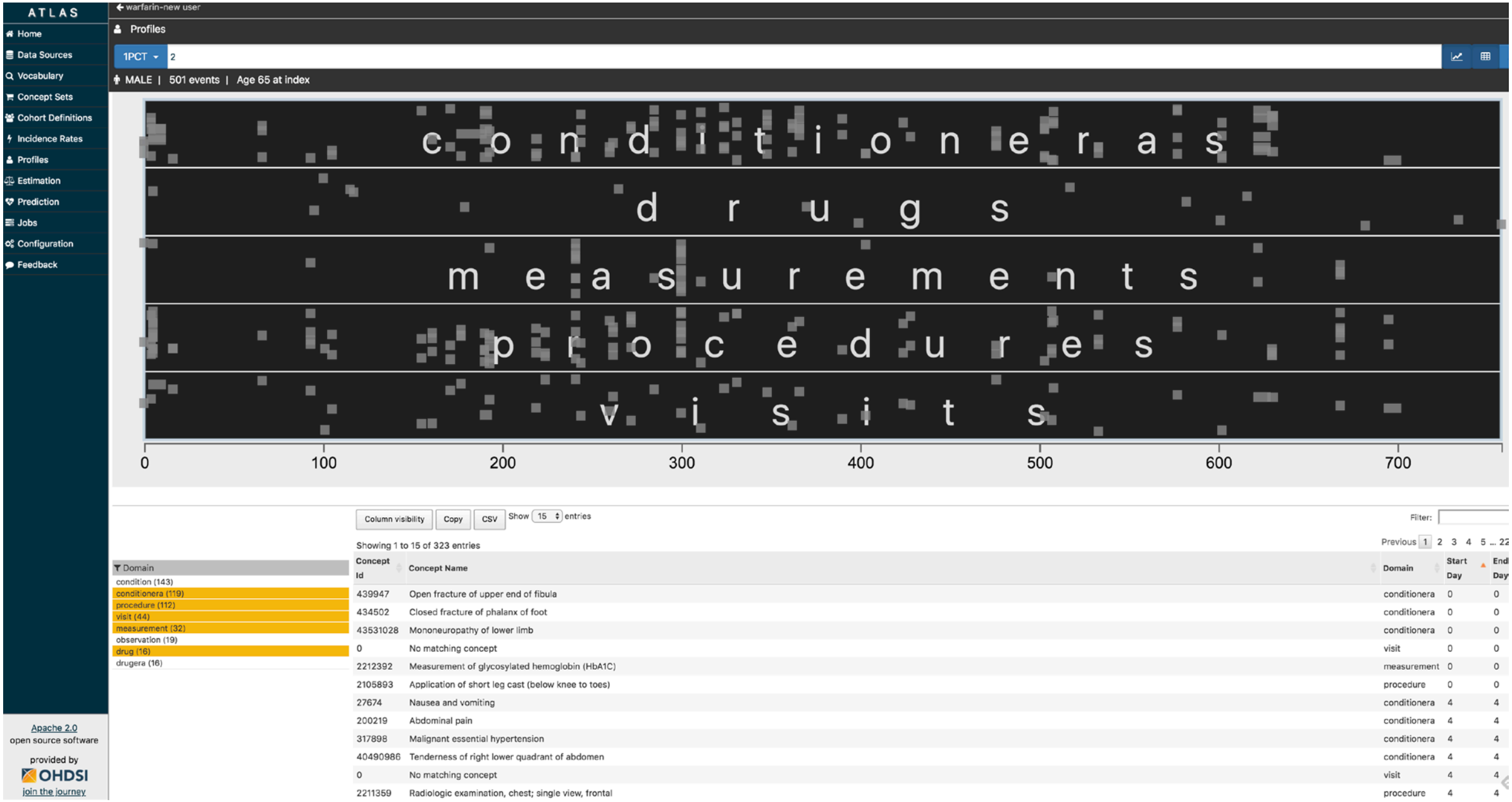
```
                Ethnicity  n proportion
1:     Hispanic or Latino  5 0.05208333
2: Not Hispanic or Latino 91 0.94791667
```

PatientExploreR

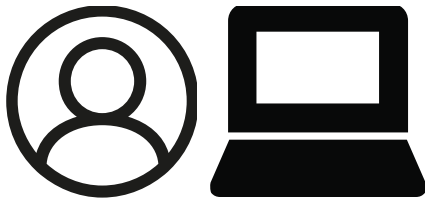
dynamic visualization of clinical history in OMOP format

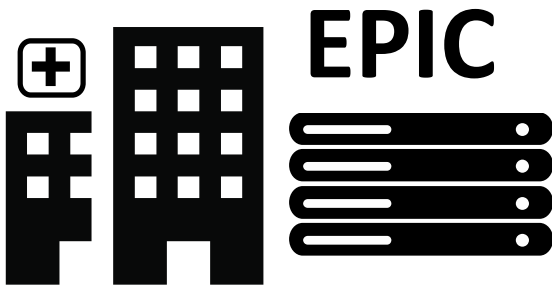
Glicksberg et al. (in revision)

No flexible application exists



Goals





PatientExploreR | Home | Patient Finder | Overall Report | Encounter Timeline | Data Explorer | More

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

glicksbergb

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 futher 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 ecnounter 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. Cateogrial 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 simulaneous and linked exploration of multiple variables.

Who We Are

Butte Lab, Institute for Computational Health Sciences, UCSF

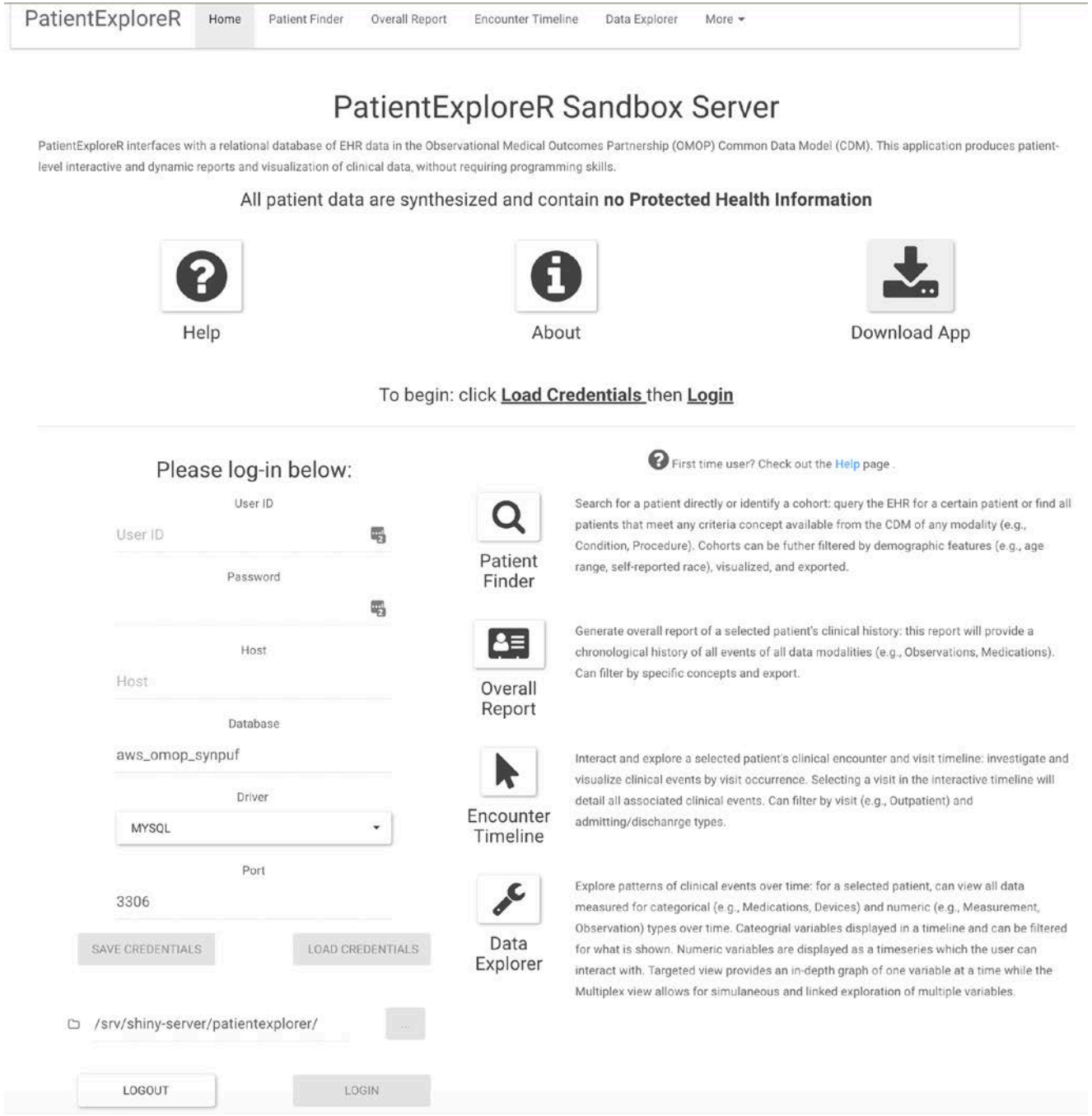
Contact & Lab Logo/Dscription

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.

Search Mode:

- ○ Search by Patient
- ◉ Search by Criteria

Criteria (select from table):

Select Domain: CONDITION

Select Vocabulary: ICD10CM

Select Concept Class: 9 ITEMS SELECTED

SELECT ALL | NONE

Search: K51

concept_code	concept_name	domain_id	vocabulary_id	concept_class_id
K51.4	Inflammatory polyps of colon	Condition	ICD10CM	4-char nonbill code
K51.414	Inflammatory polyps of colon with abscess	Condition	ICD10CM	6-char billing code
K51.41	Inflammatory polyps of colon with complications	Condition	ICD10CM	5-char nonbill code
K51.413	Inflammatory polyps of colon with fistula	Condition	ICD10CM	6-char billing code
K51.412	Inflammatory polyps of colon with intestinal obstruction	Condition	ICD10CM	6-char billing code

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

Previous 1 2 3 4 5 ... 13 Next

Selected Criteria:

vocabulary	term
ICD10CM	K51.4
ICD10CM	K51.414
ICD10CM	K51.41
ICD10CM	K51.413
ICD10CM	K51.412

Showing 1 to 5 of 64 entries

Previous 1 2 3 4 5 ... 13 Next

REMOVE ITEM

RESET SEARCH

Search Type:

- ◉ or
- ○ and

Search Strategy:

- ◉ Mapped
- ○ Direct

SEARCH BY CRITERIA

Filter Cohort:

Age

22

22 31 40 49

RESET FILTERS

25
20
15
10
5
0
20 30 40

Filter Cohort:

Age

22 74 109

22 31 40 49 58 67 76 85 94 103109

Gender

MALE

Status

ALIVE

Race

3 ITEMS SELECTED

Ethnicity

2 ITEMS SELECTED

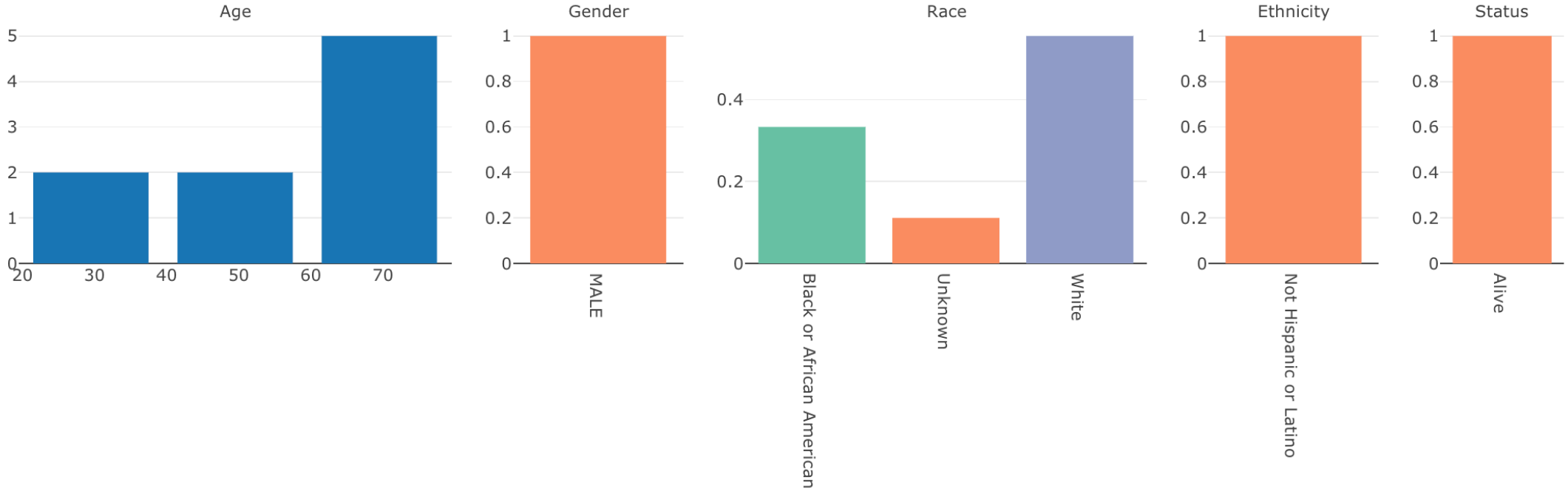
RESET FILTERS

EXPORT COHORT

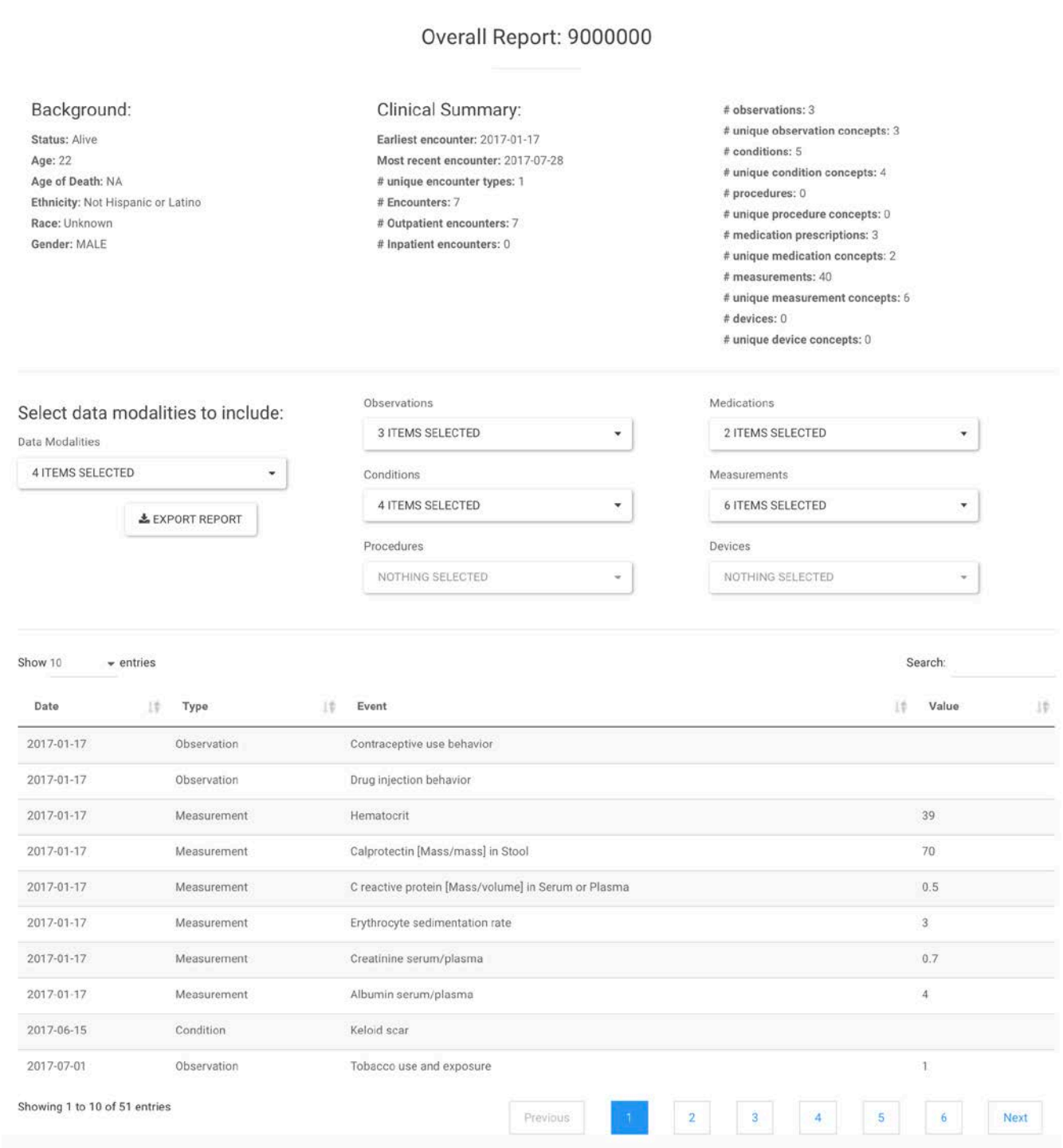
HIDE PLOTS

Selected Patient ID:

SEARCH



Automatically generated clinical history



Encounters Timeline: 9000000

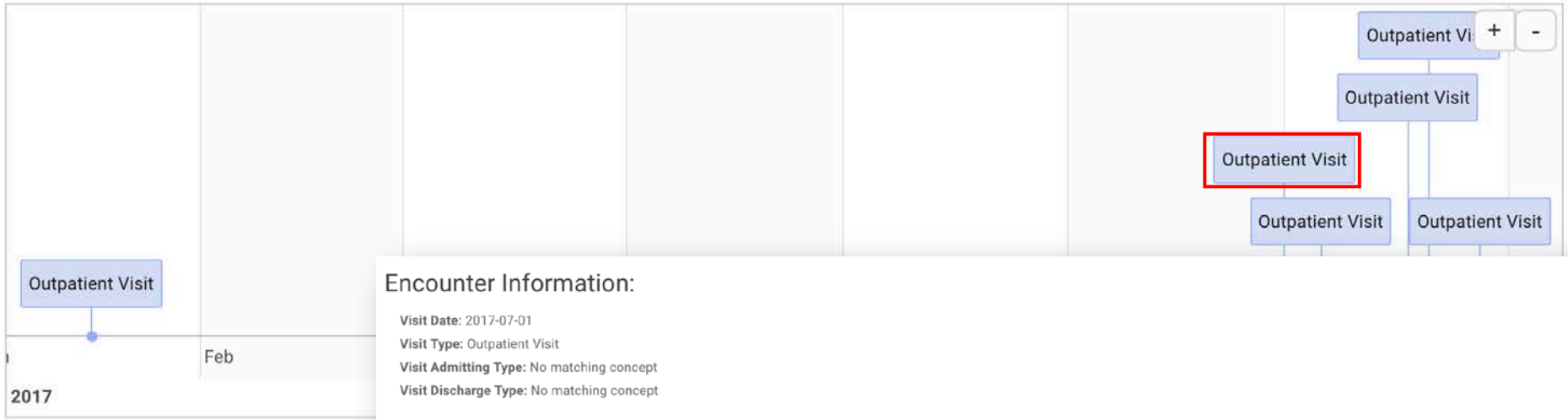
Plot Encounters:

- (●) None
- ( ) Visit Types
- ( ) Admitting Concepts
- ( ) Discharge Concepts

Visit Types: OUTPATIENT VISIT

Admitting Concept Type: NO MATCHING CONCEPT

Discharge Concept Type: NO MATCHING CONCEPT



FIT ALL ENCOUNTERS | FOCUS PAST YEA

Encounter Information:

Visit Date: 2017-07-01
Visit Type: Outpatient Visit
Visit Admitting Type: No matching concept
Visit Discharge Type: No matching concept

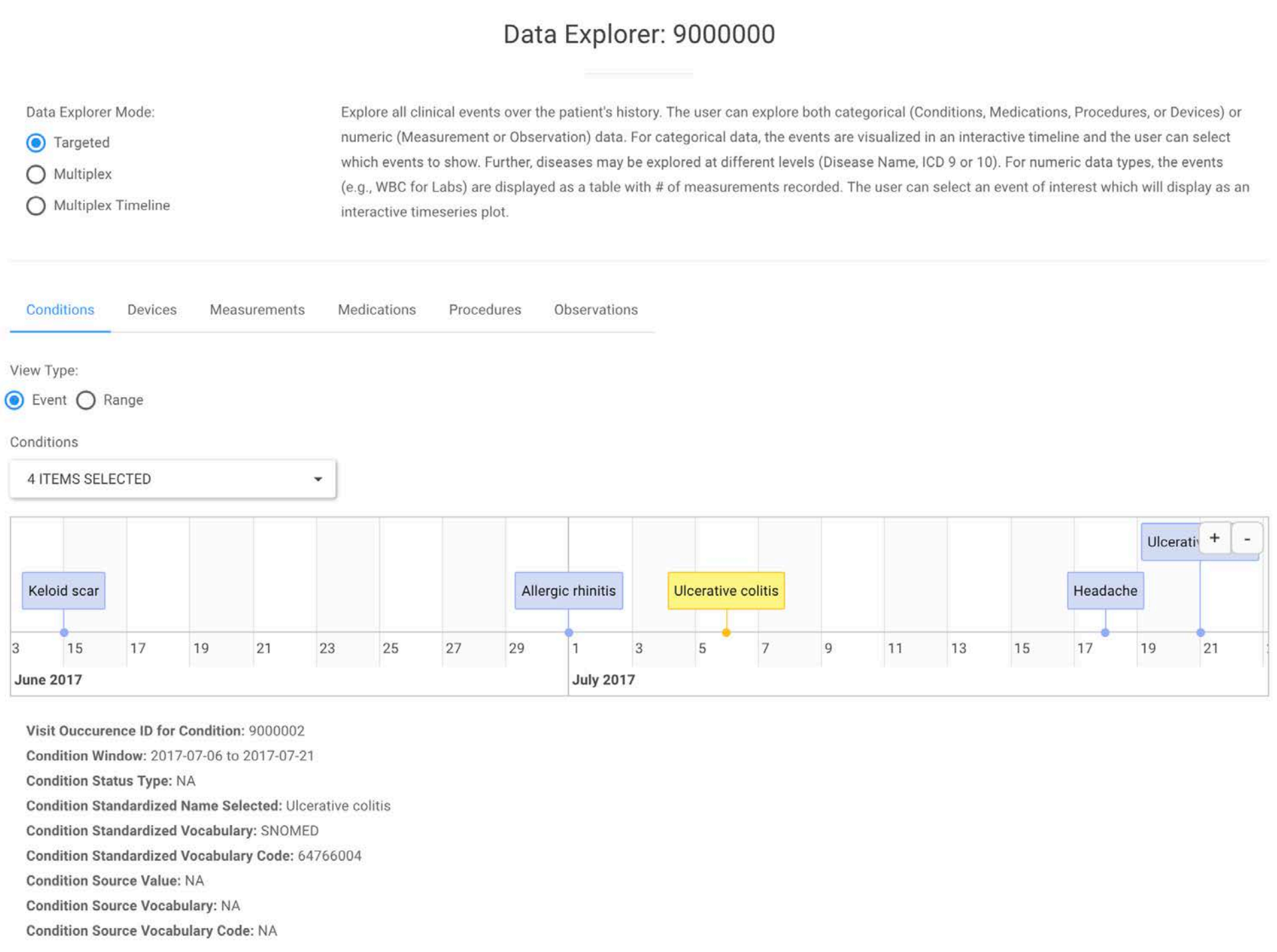
Conditions | Devices | Measurements | Medications | Observations | Procedures

Show 10 entries Search:

condition_concept_name	condition_type	condition_status_type	condition_concept_vocabulary	condition_concept_code	condition_source_vocabulary	condition_source_code	condition_start_date	condition_end_date
Allergic rhinitis	Primary Condition		SNOMED	61582004			2017-07-01	2017-07-10

Showing 1 to 1 of 1 entries Previous 1 Next

Explore Trends in Data/ Outcomes (targeted)



Explore Trends in Data/ Outcomes (numeric; targeted)

Conditions | Devices | Measurements | Medications | Procedures | Observations

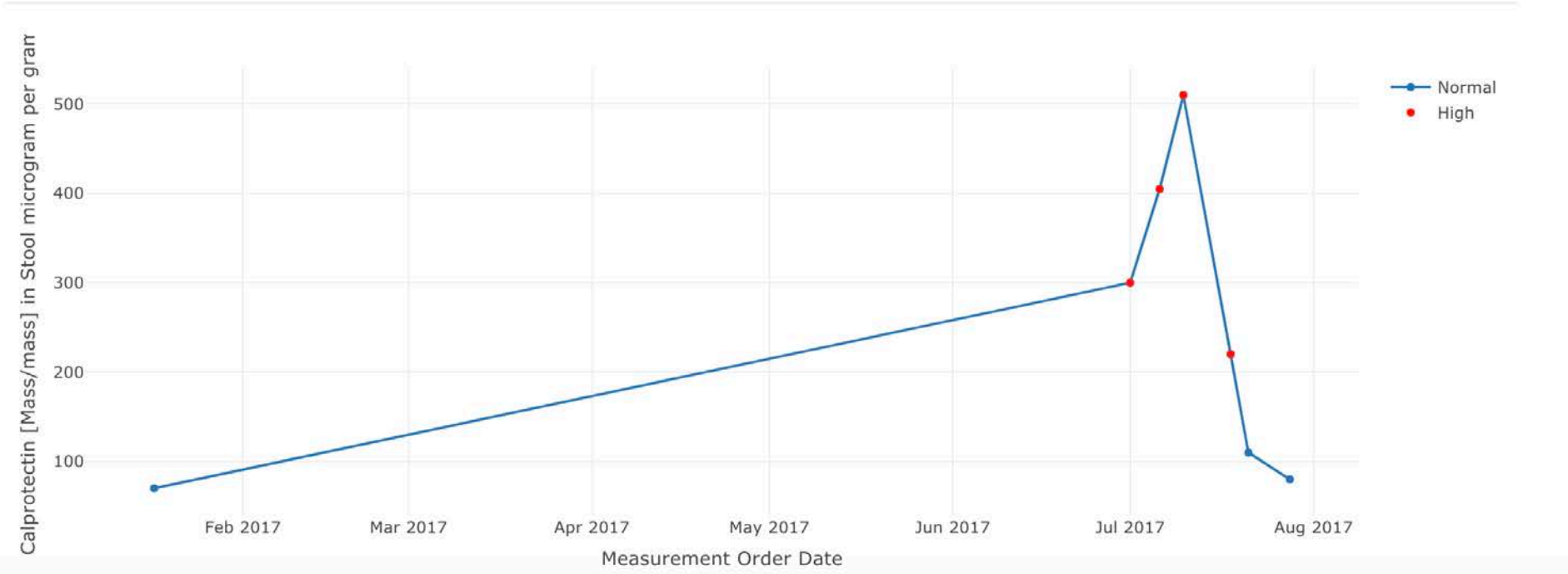
Show 5 entries

Search:

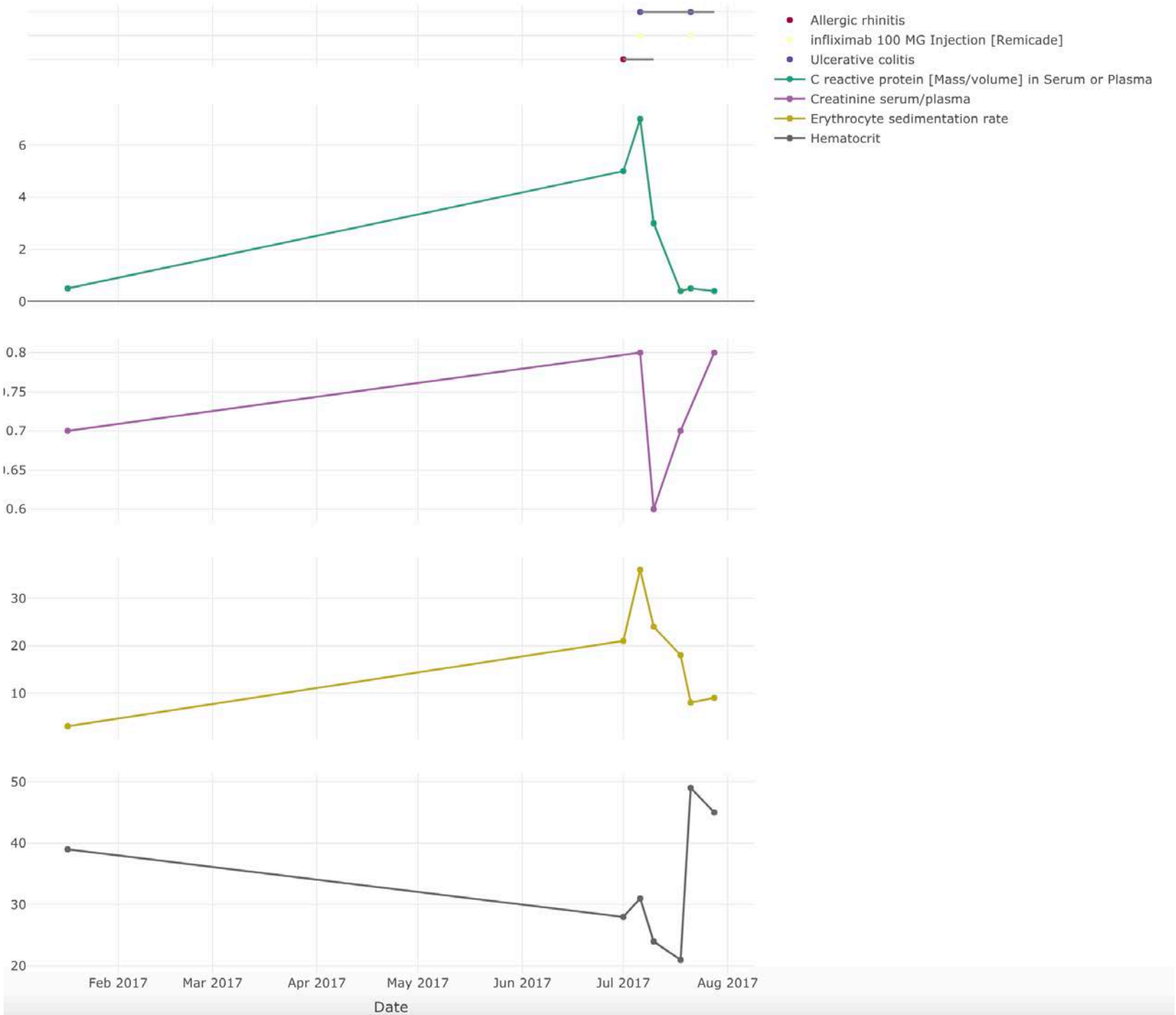
Measurement Concept Name	N
Albumin serum/plasma	7
Calprotectin [Mass/mass] in Stool	7
C reactive protein [Mass/volume] in Serum or Plasma	7
Erythrocyte sedimentation rate	7
Hematocrit	7

Showing 1 to 5 of 6 entries

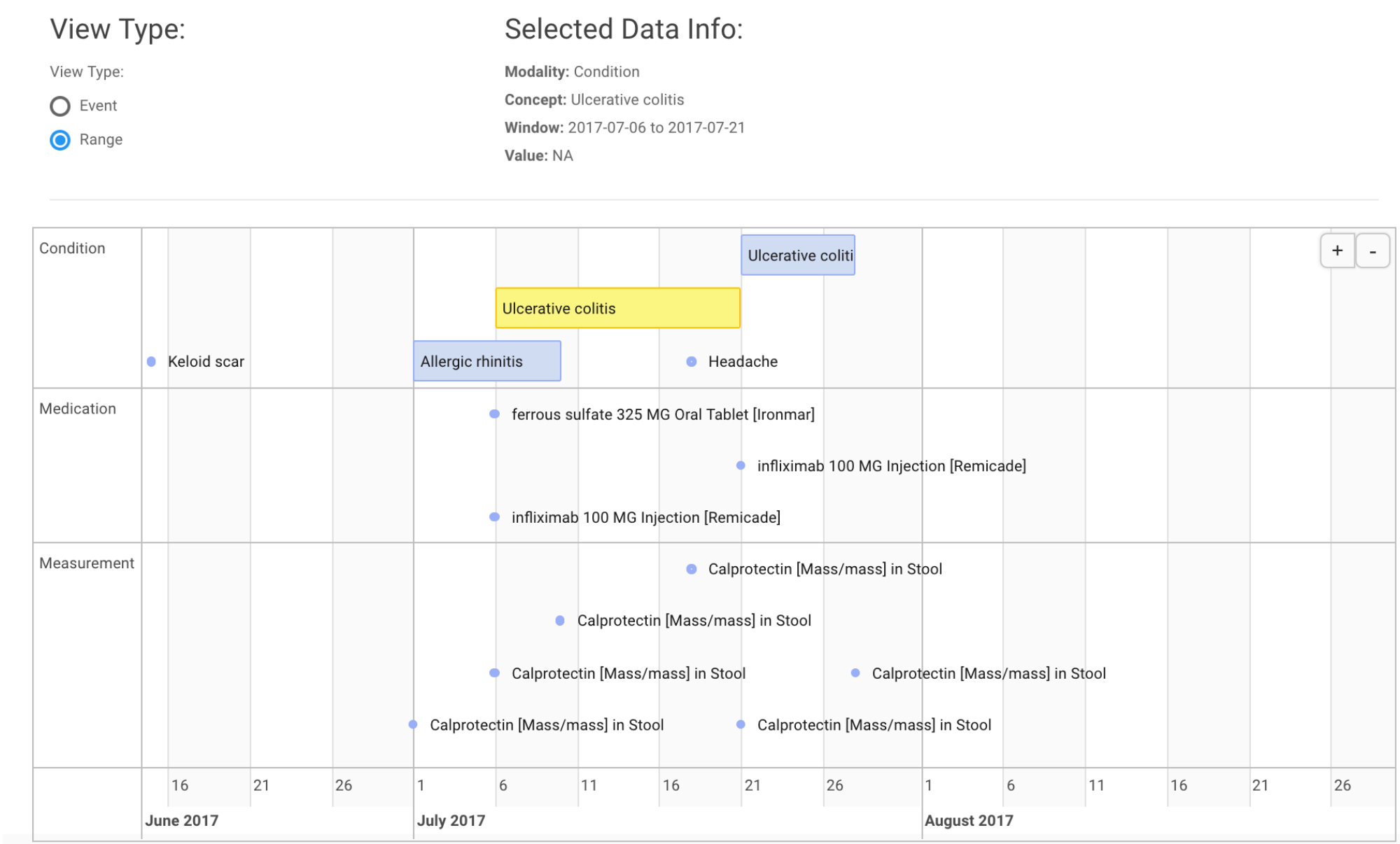
Previous 1 2 Next



Explore Trends in Data/ Outcomes (multiplex)



Explore Trends in Data/ Outcomes (multiplex timeline)



How might these tools enable AI-based EHR research?

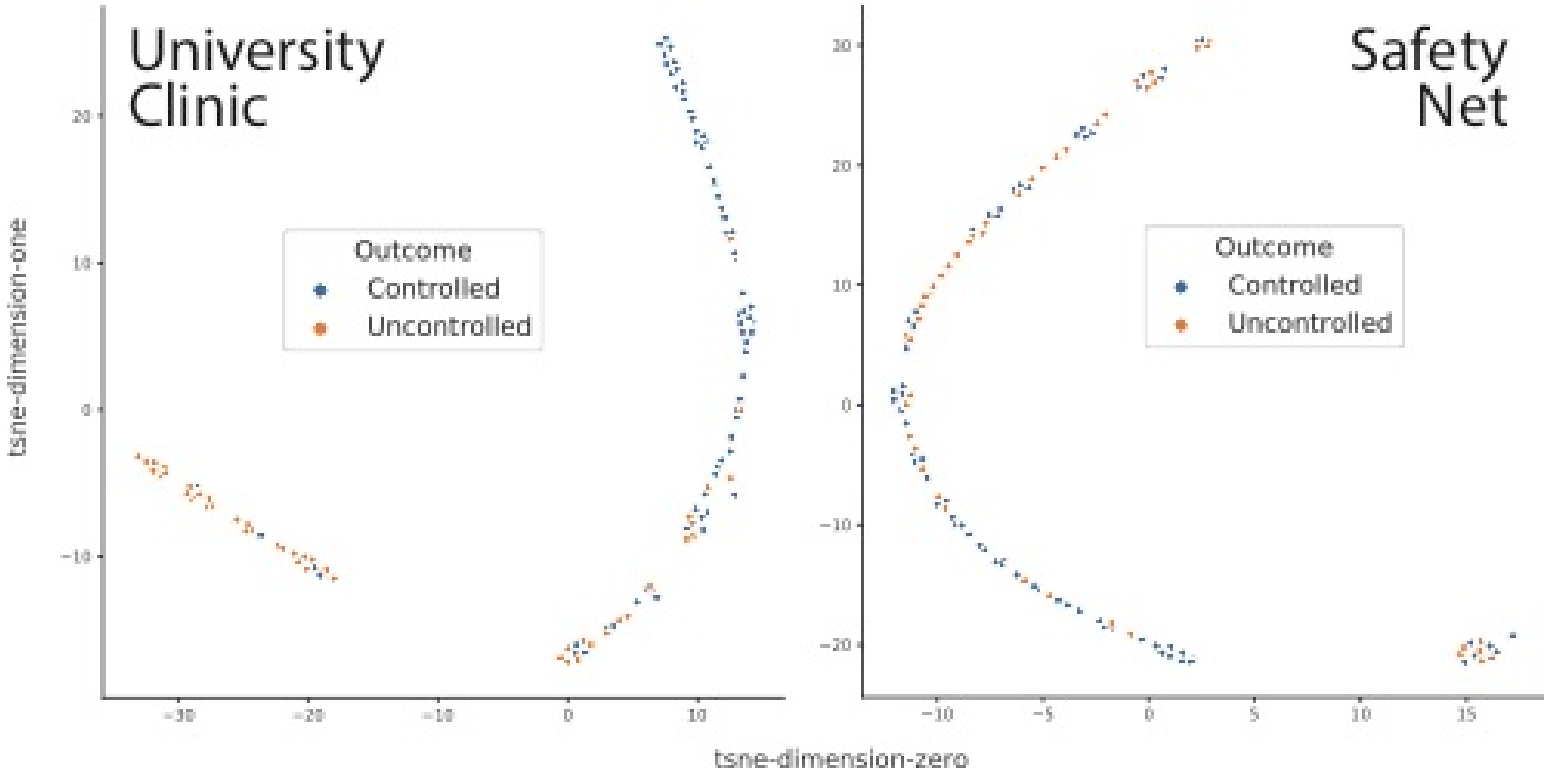
How well can we predict…

- Risk for disease
- Disease onset
- Symptom severity
- Treatment response
- Medication adverse events
- Ideal dose of medication
- Symptom flares
- Length of stay in hospital

Time Aggregation and Model Interpretation for Deep Multivariate Longitudinal Patient Outcome Forecasting Systems in Chronic Ambulatory Care

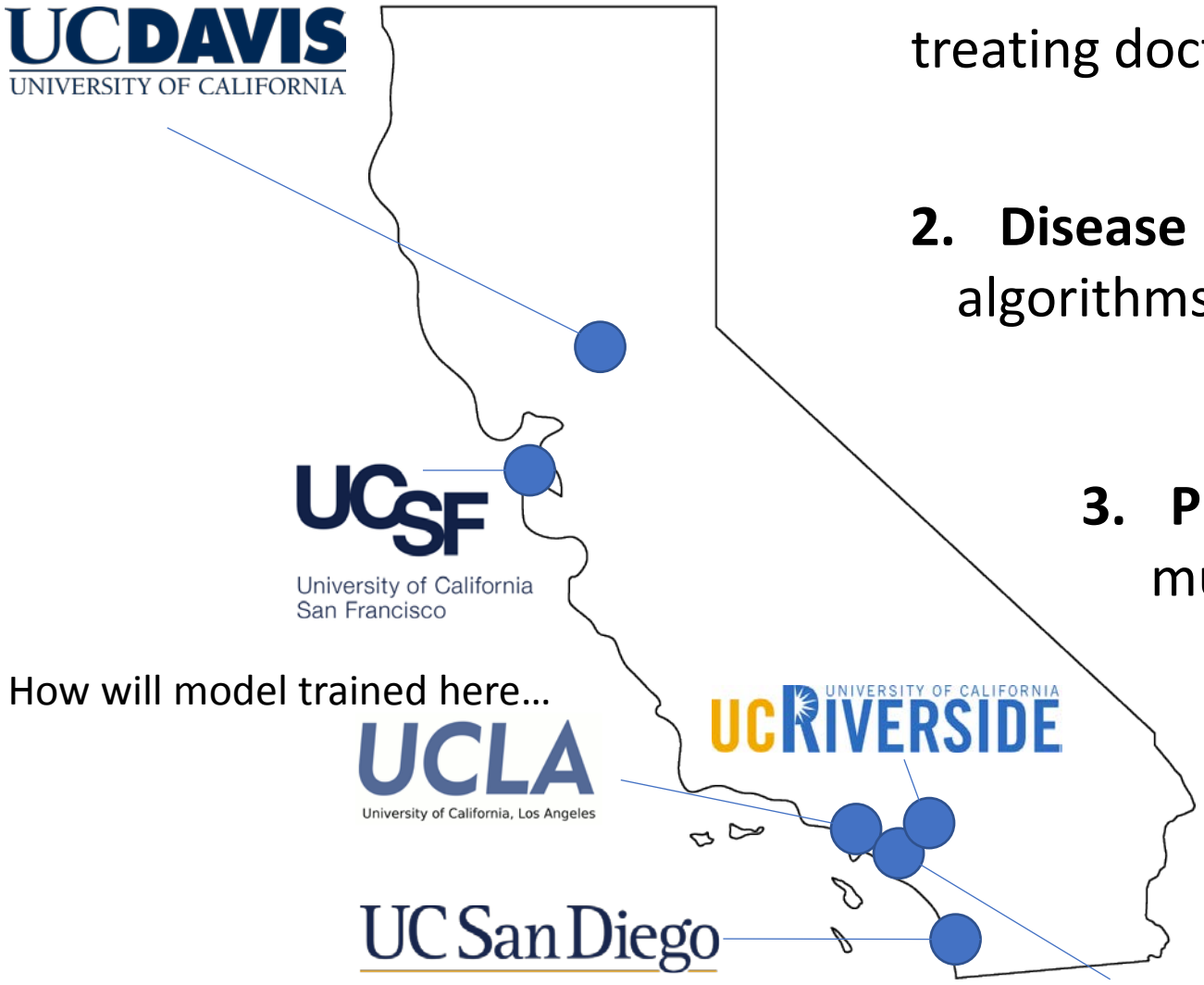
Beau Norgeot[1] Dmytro Lituiev[1] Benjamin S. Glicksberg[1] Atul J. Butte[1*]

[1] Bakar Computational Health Sciences Institute, University of California, San Francisco
*atul.butte@ucsf.edu



Beau Norgeot, MS

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

Butte Lab

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

UCSF Bakar Computational Health Sciences Institute

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

Icahn School of Medicine at Mount Sinai — Graduate School of Biomedical Sciences

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

UCSF Information Technology

Dana Ludwig, MD
Remi Frazier
Nelson Lee
Rick Larsen

Columbia University Medical Center
Discover. Educate. Care. Lead.

Nicholas Giangreco
Phyllis Thangaraj
Nicholas Tatonetti, PhD

OHDSI
Observational Health Data Sciences and Informatics

Community and Developers

R