Observational Health Data Sciences and Informatics (OHDSI)

George Hripcsak, MD, MS
Columbia University Medical Center
NewYork-Presbyterian Hospital

Seattle Symposium on Health Care Data Analytics

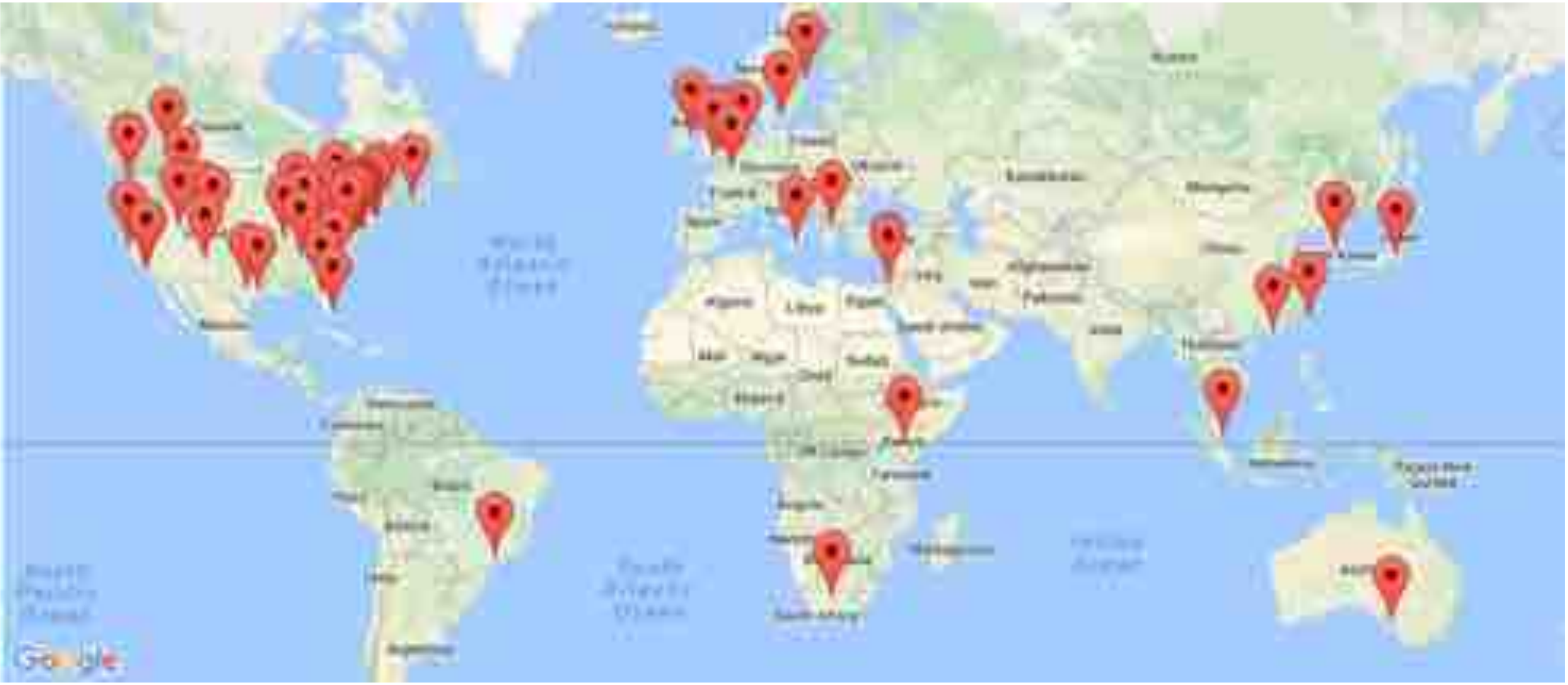
Observational Health Data Sciences and Informatics (OHDSI, as "Odyssey")

A multi-stakeholder, interdisciplinary, international collaborative with a coordinating center at Columbia University

Mission: To improve health, by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care

Aiming for 1,000,000,000 patient data network

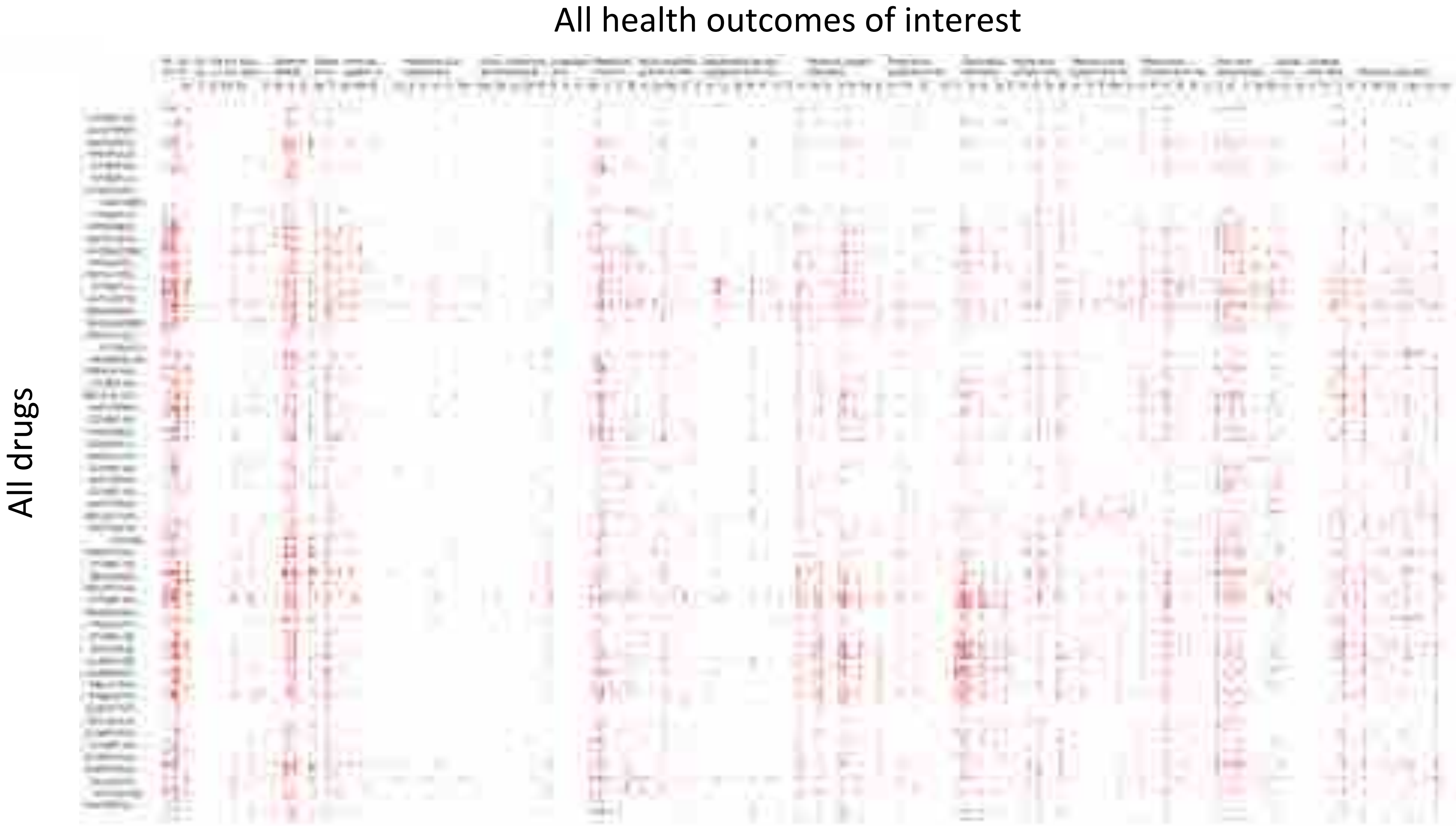
OHDSI's global research community



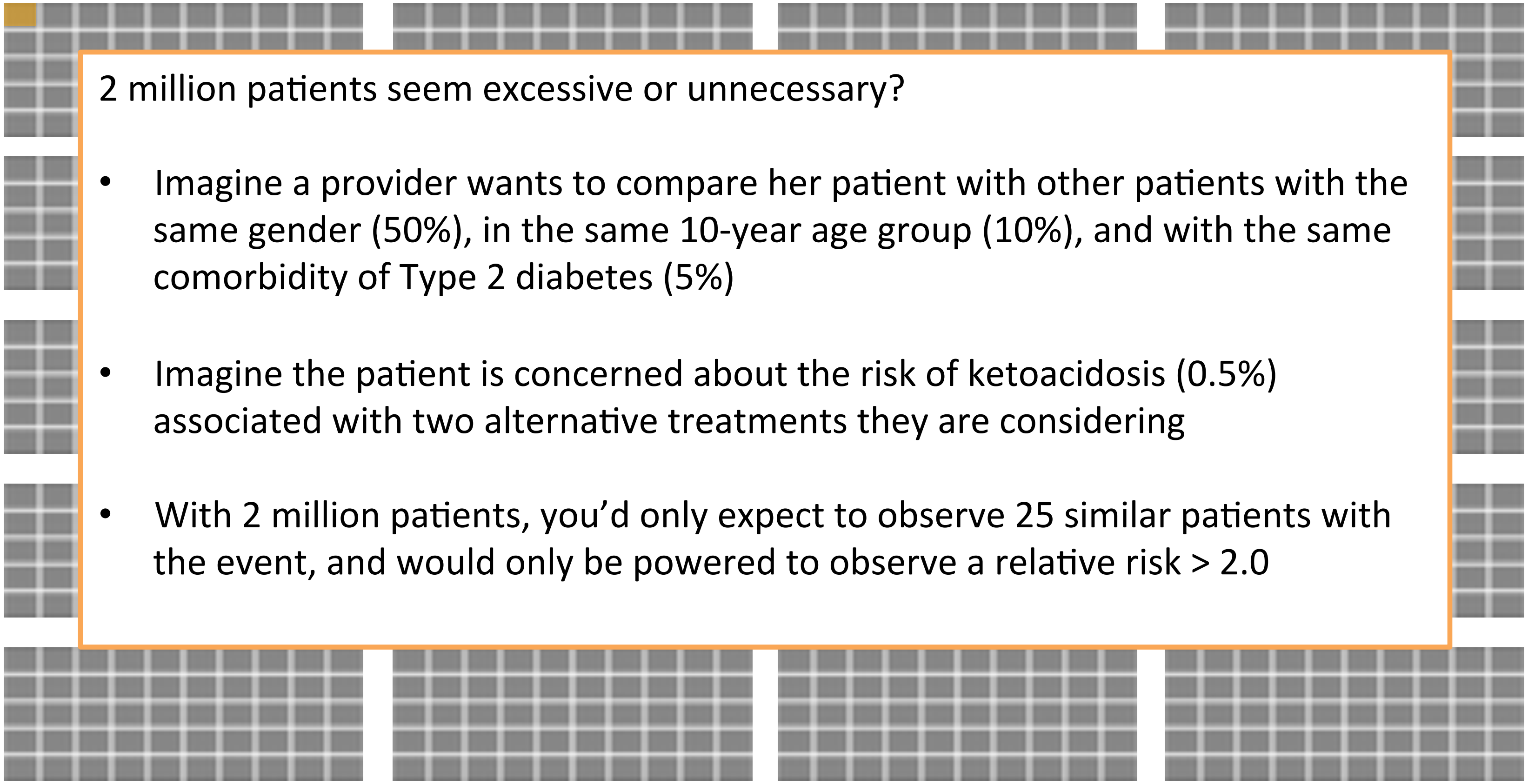
- >140 collaborators from 20 different countries
- Experts in informatics, statistics, epidemiology, clinical sciences
- Active participation from academia, government, industry, providers
- Currently 600 million patient records in 52 databases

http://ohdsi.org/who-we-are/collaborators/

Why large-scale analysis is needed in healthcare



Patient-level predictions for personalized evidence requires big data



2 million patients seem excessive or unnecessary?

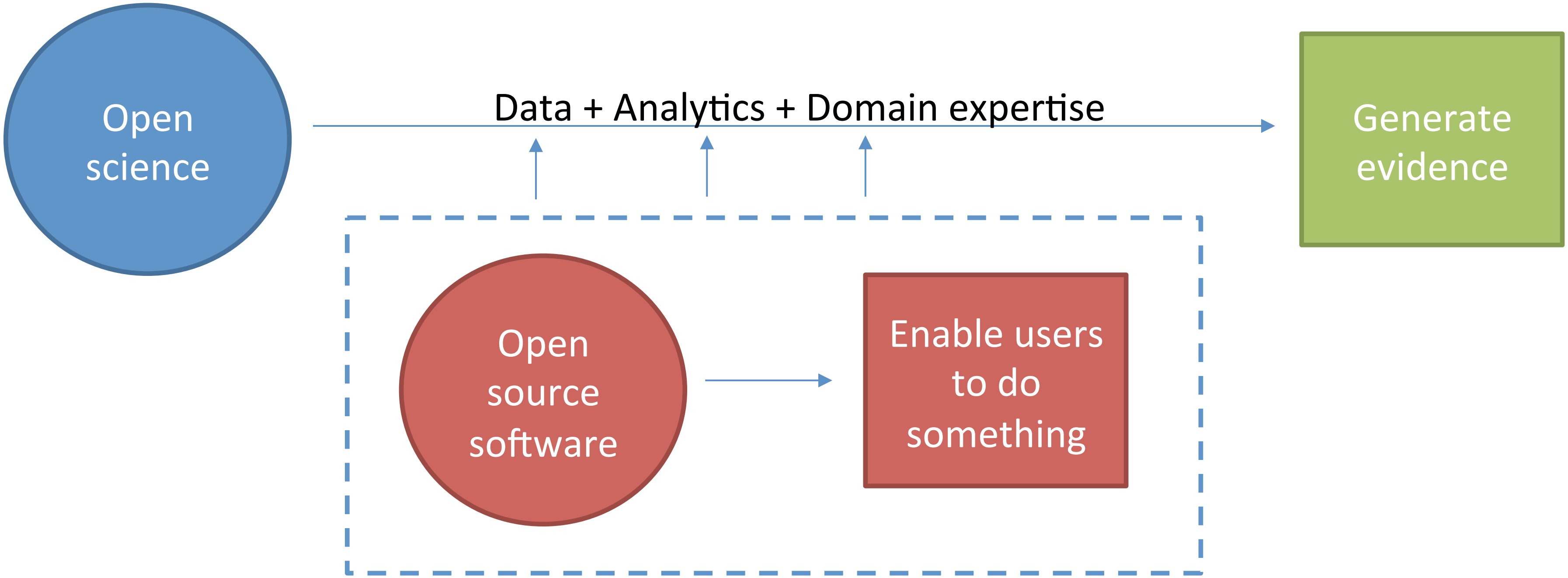
- Imagine a provider wants to compare her patient with other patients with the same gender (50%), in the same 10-year age group (10%), and with the same comorbidity of Type 2 diabetes (5%)
- Imagine the patient is concerned about the risk of ketoacidosis (0.5%) associated with two alternative treatments they are considering
- With 2 million patients, you'd only expect to observe 25 similar patients with the event, and would only be powered to observe a relative risk > 2.0

Aggregated data across a health system of 1,000 providers may contain 2,000,000 patients

Evidence OHDSI seeks to generate from observational data

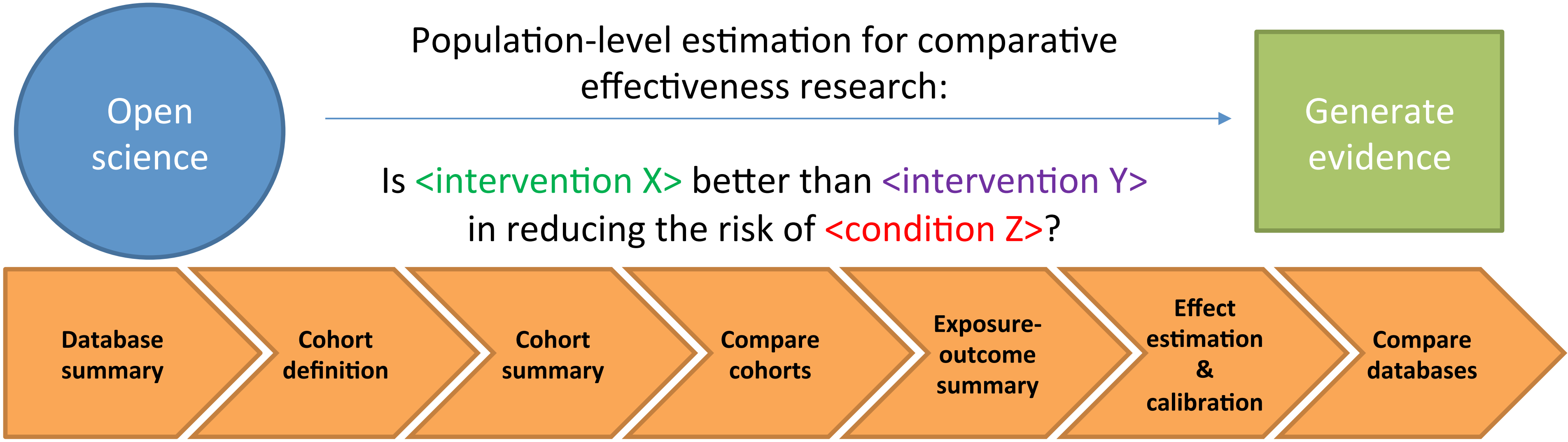
- **Clinical characterization**
  - Natural history: Who has diabetes, and who takes metformin?
  - Quality improvement: What proportion of patients with diabetes experience complications?
- **Population-level estimation**
  - Safety surveillance: Does metformin cause lactic acidosis?
  - Comparative effectiveness: Does metformin cause lactic acidosis more than glyburide?
- **Patient-level prediction**
  - Precision medicine: Given everything you know about me, if I take metformin, what is the chance I will get lactic acidosis?
  - Disease interception: Given everything you know about me, what is the chance I will develop diabetes?

OHDSI’s approach to open science


Open science
Data + Analytics + Domain expertise
Generate evidence
Open source software
Enable users to do something

- Open science is about sharing the journey to evidence generation
- Open-source software can be part of the journey, but it’s not a final destination
- Open processes can enhance the journey through improved reproducibility of research and expanded adoption of scientific best practices

Standardizing workflows to enable transparent, reproducible research



Open science

Population-level estimation for comparative effectiveness research:

Is <intervention X> better than <intervention Y> in reducing the risk of <condition Z>?

Generate evidence

Database summary → Cohort definition → Cohort summary → Compare cohorts → Exposure-outcome summary → Effect estimation & calibration → Compare databases

Defined inputs:

- Target exposure
- Comparator group
- Outcome
- Time-at-risk
- Model specification

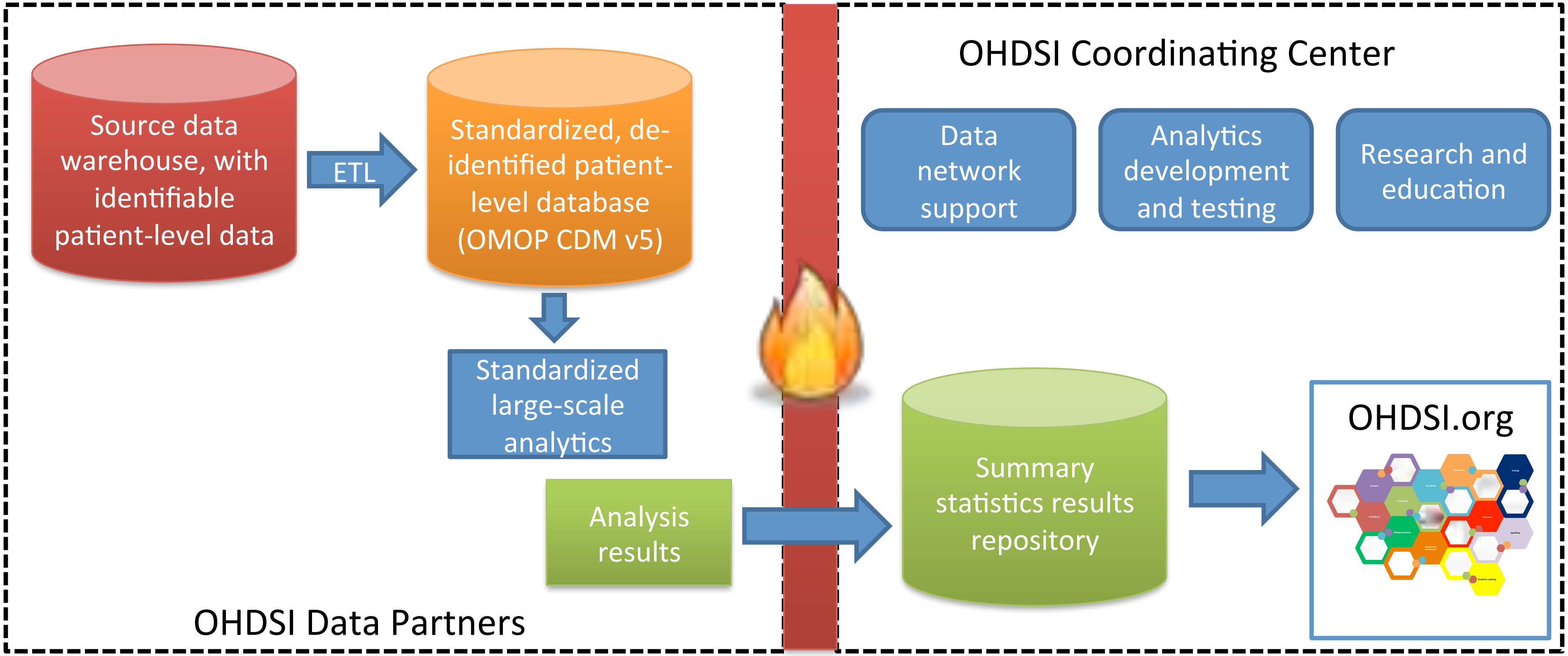
Consistent outputs:

- analysis specifications for transparency and reproducibility (protocol + source code)
- only aggregate summary statistics (no patient-level data)
- model diagnostics to evaluate accuracy
- results as evidence to be disseminated
  - static for reporting (e.g. via publication)
  - interactive for exploration (e.g. via app)

OHDSI Distinguishing Features

- International effort (size & coverage)
  - – 43 sources terminologies from around the world
- Open science (depth)
  - – Infrastructure serves the science
  - – Stack: Terminology, CDM, ETL, QA, Visualization, Novel analytic methods, Clinical research
- Full information model

How OHDSI Works



OHDSI Coordinating Center

Data network support

Analytics development and testing

Research and education

Source data warehouse, with identifiable patient-level data

ETL

Standardized, de-identified patient-level database (OMOP CDM v5)

Standardized large-scale analytics

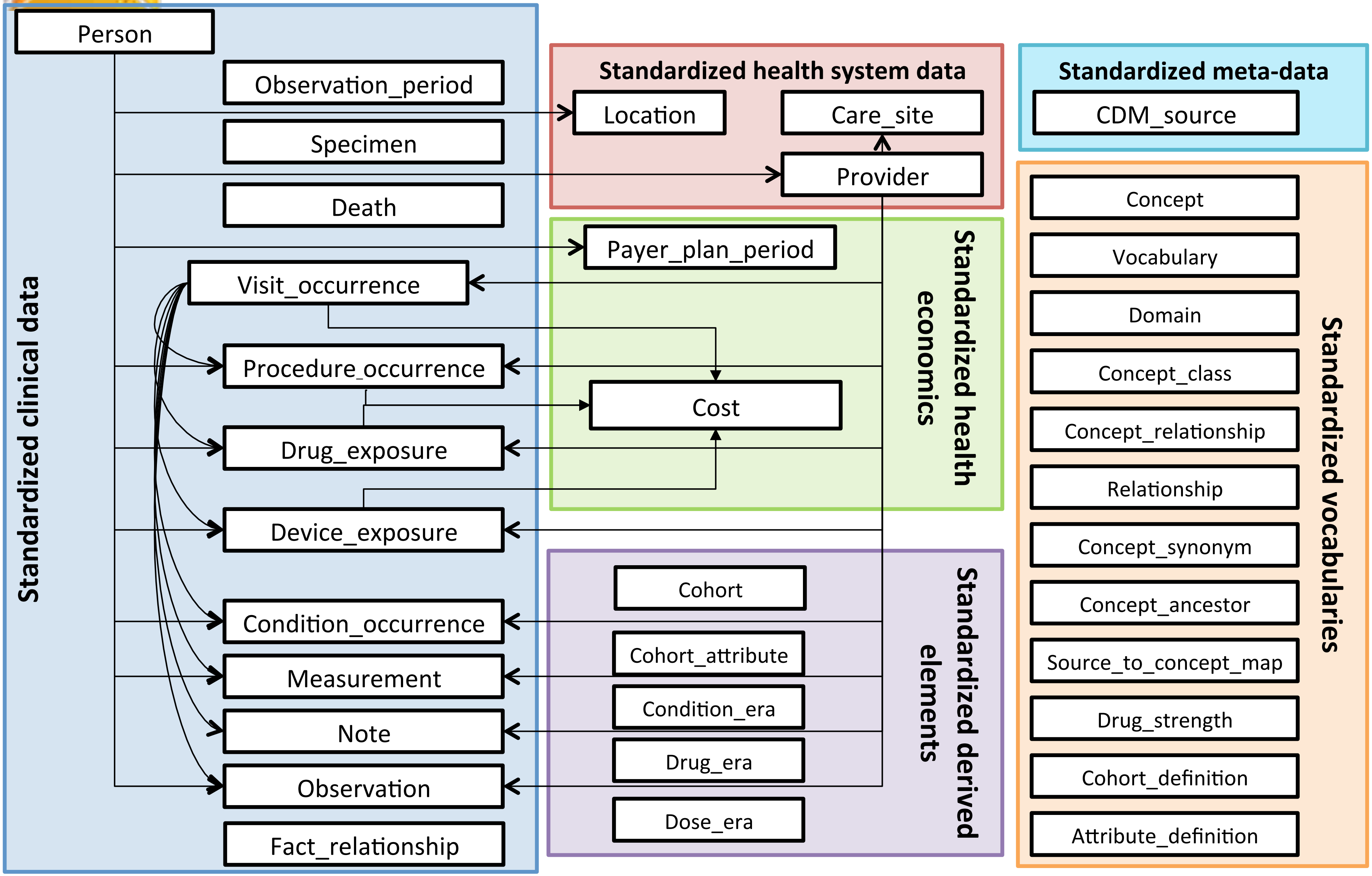
Analysis results

Summary statistics results repository

OHDSI.org

OHDSI Data Partners

Deep information model
OMOP CDM v5.0.1



Standardized clinical data

Person

Observation_period

Specimen

Death

Visit_occurrence

Procedure_occurrence

Drug_exposure

Device_exposure

Condition_occurrence

Measurement

Note

Observation

Fact_relationship

Standardized health system data

Location

Care_site

Provider

Standardized health economics

Payer_plan_period

Cost

Standardized derived elements

Cohort

Cohort_attribute

Condition_era

Drug_era

Dose_era

Standardized meta-data

CDM_source

Standardized vocabularies

Concept

Vocabulary

Domain

Concept_class

Concept_relationship

Relationship

Concept_synonym

Concept_ancestor

Source_to_concept_map

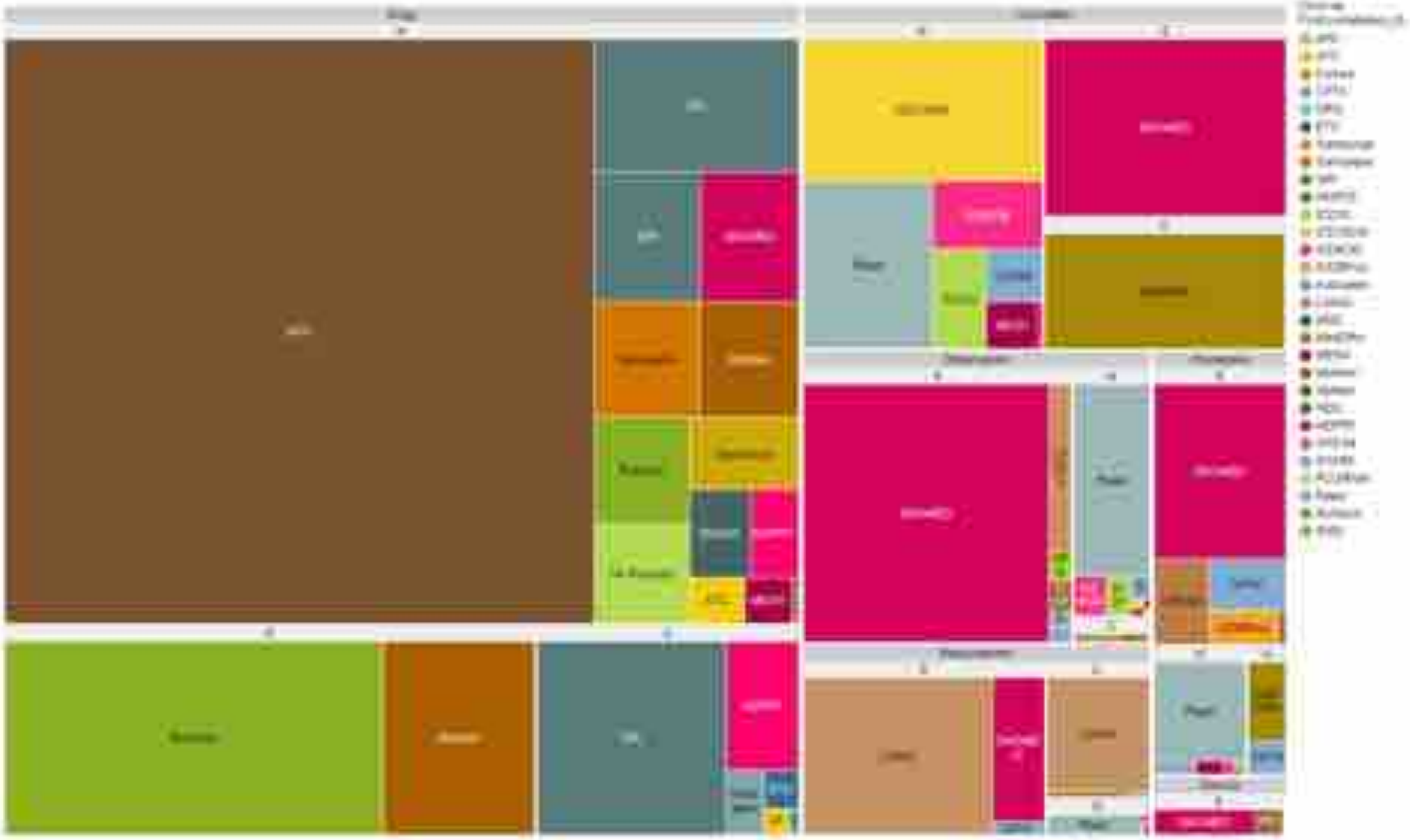
Drug_strength

Cohort_definition

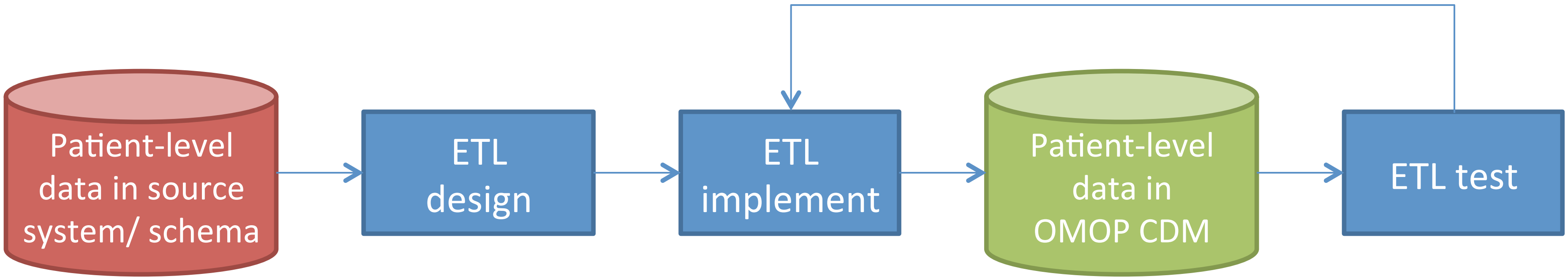
Attribute_definition

Extensive vocabularies

Breakdown of OHDSI concepts by domain, standard state, and vocabulary



Preparing your data for analysis



Patient-level data in source system/ schema

ETL design

ETL implement

Patient-level data in OMOP CDM

ETL test

OHDSI tools built to help

WhiteRabbit: profile your source data

RabbitInAHat: map your source structure to CDM tables and fields

ATHENA: standardized vocabularies for all CDM domains

Usagi: map your source codes to CDM vocabulary

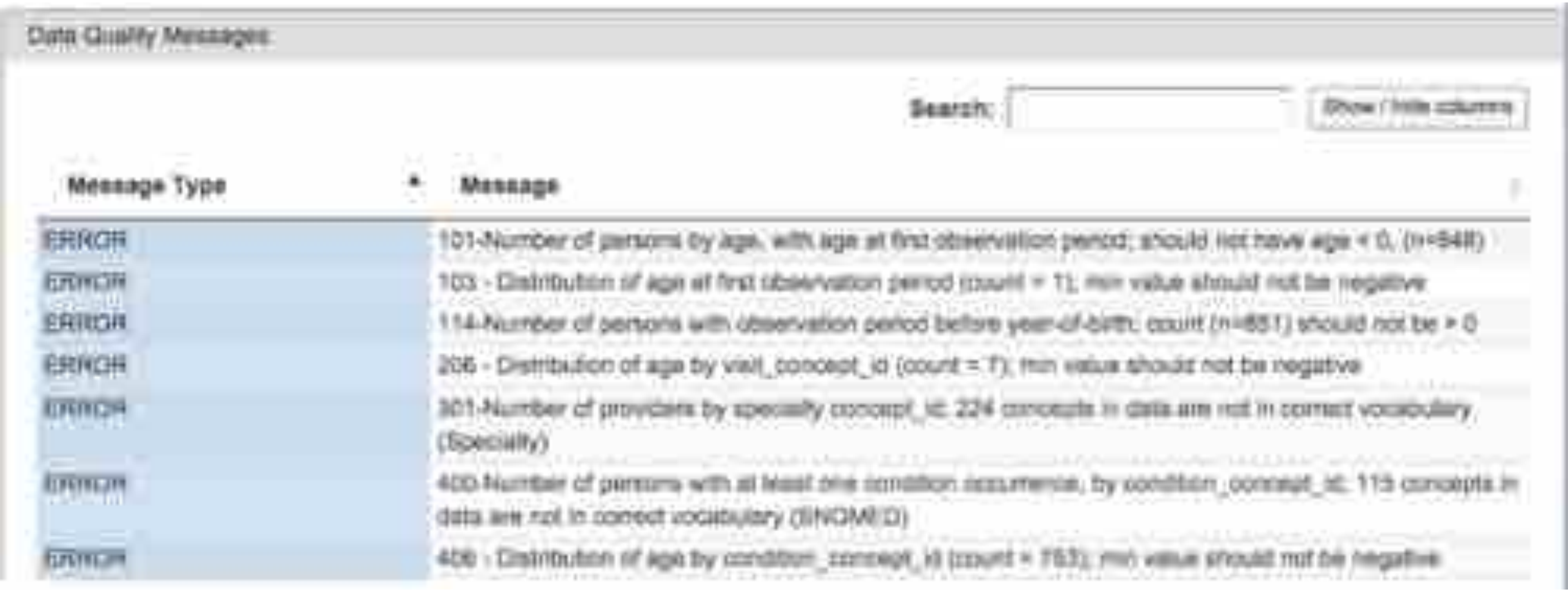
CDM: DDL, index, constraints for Oracle, SQL Server, PostgresQL; Vocabulary tables with loading scripts

ACHILLES: profile your CDM data; review data quality assessment; explore population-level summaries

OHDSI Forums: Public discussions for OMOP CDM Implementers/developers

http://github.com/OHDSI

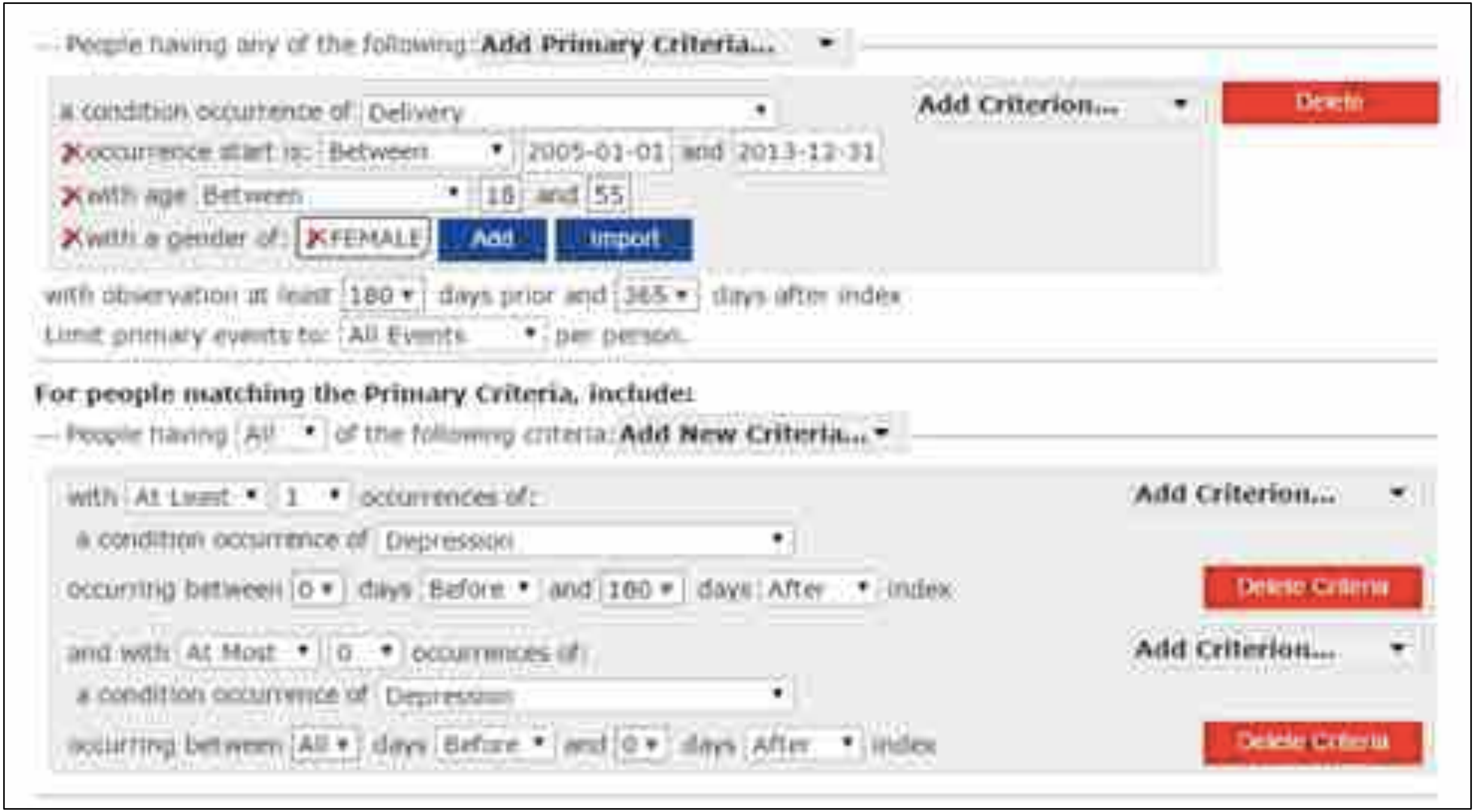
ACHILLES Heel Data Validation



Data Quality Messages

Message Type	Message
ERROR	101-Number of persons by age, with age at first observation period; should not have age < 0, (n=948)
ERROR	103 - Distribution of age at first observation period (count = 1); min value should not be negative
ERROR	114-Number of persons with observation period before year-of-birth; count (n=651) should not be > 0
ERROR	206 - Distribution of age by visit_concept_id (count = 7); min value should not be negative
ERROR	301-Number of providers by specialty concept_id; 224 concepts in data are not in correct vocabulary (Specialty)
ERROR	400-Number of persons with at least one condition occurrence, by condition_concept_id; 115 concepts in data are not in correct vocabulary (SNOMED)
ERROR	406 - Distribution of age by condition_concept_id (count = 153); min value should not be negative

ATLAS to build, visualize, and analyze cohorts



People having any of the following:

a condition occurrence of: Delivery

occurrence start is: Between 2005-01-01 and 2013-12-31

with age Between 18 and 55

with a gender of: FEMALE

with observation at least 180 days prior and 365 days after index

Limit primary events to: All Events per person.

For people matching the Primary Criteria, include:

People having All of the following criteria:

with At Least 1 occurrences of:

a condition occurrence of Depression

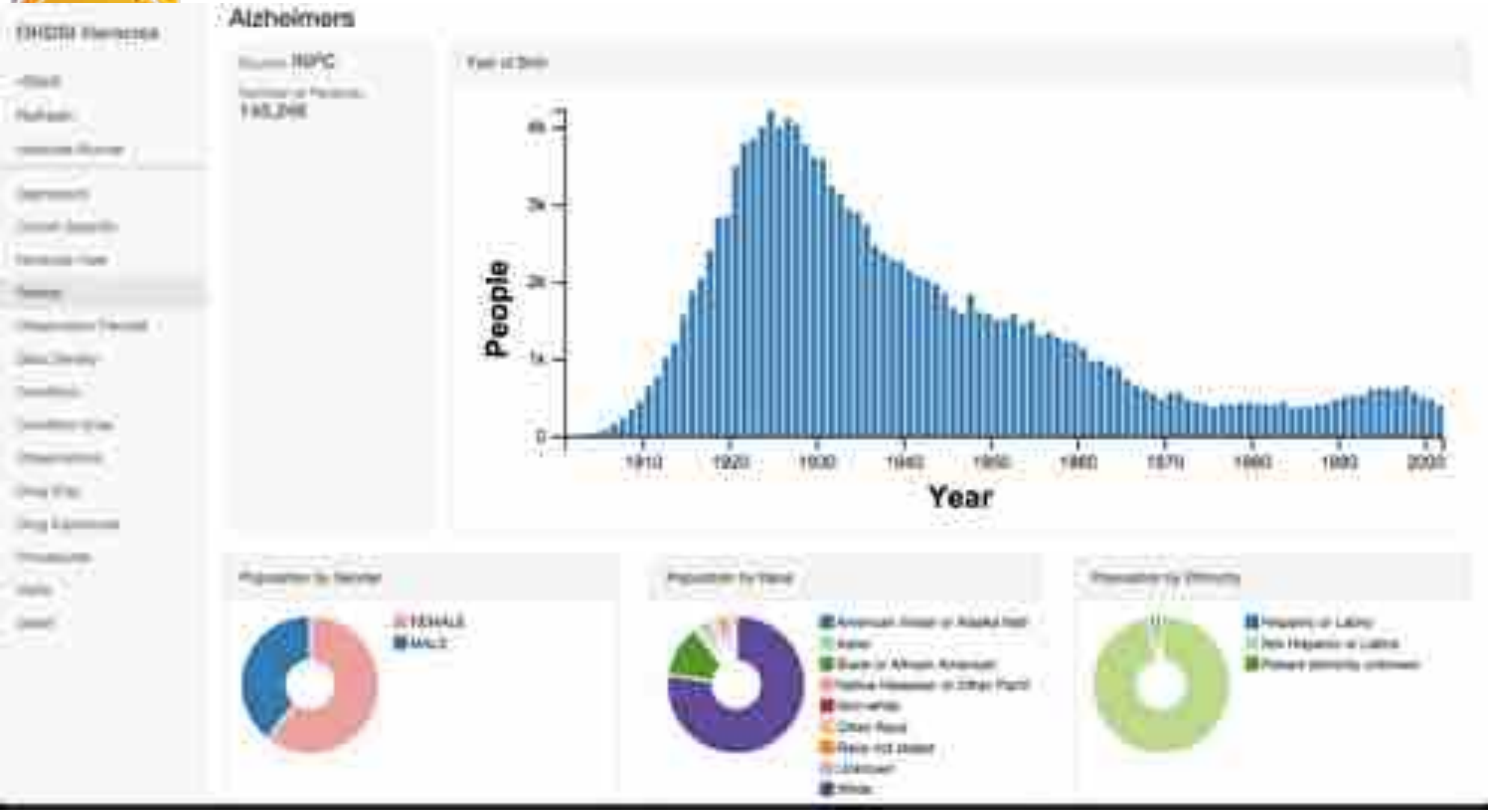
occurring between 0 days Before and 180 days After index

and with At Most 0 occurrences of:

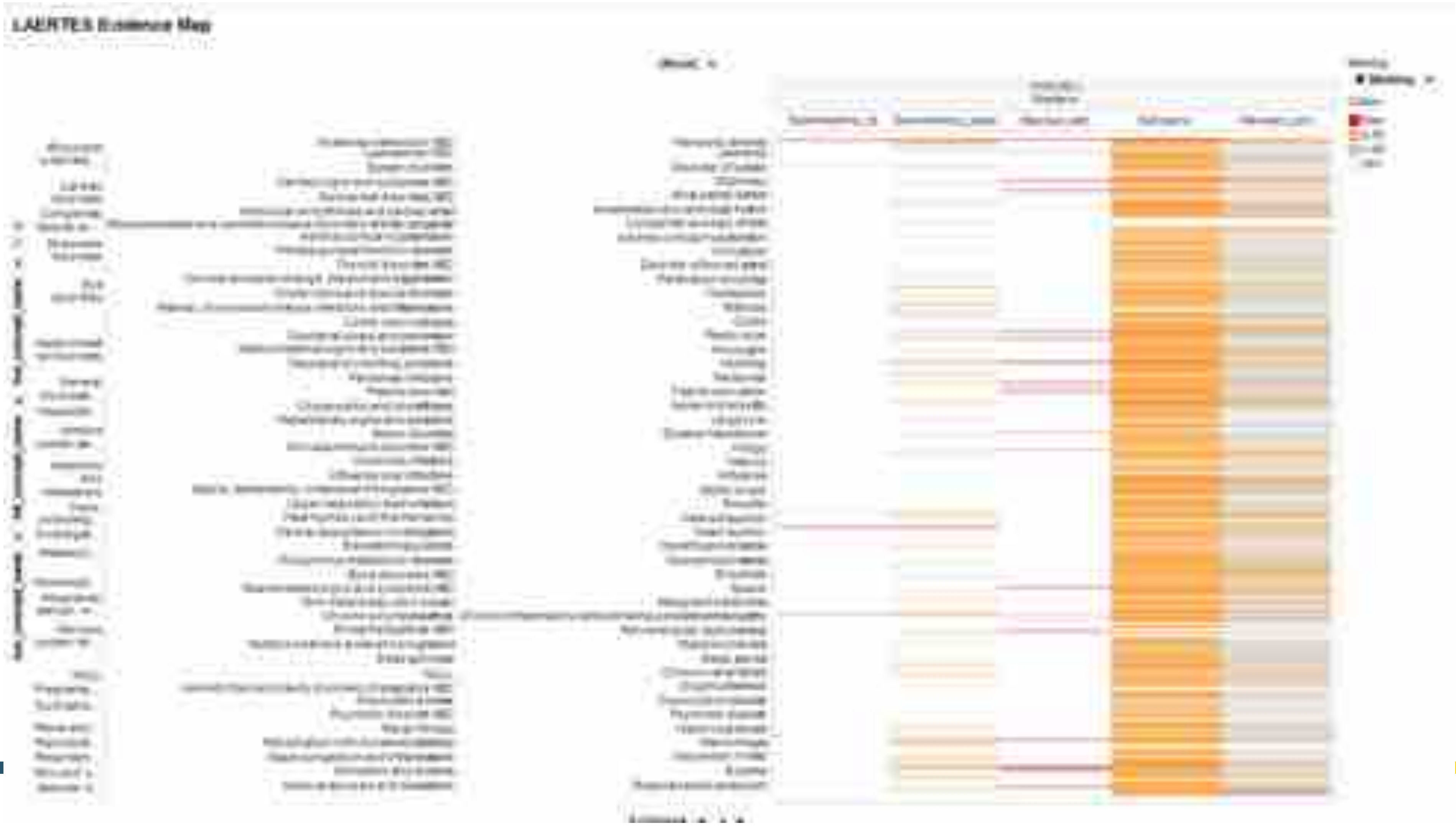
a condition occurrence of Depression

occurring between All days Before and 0 days After index

Characterize the cohorts of interest



LAERTES: Knowledge base of what we know: literature, labeling, spontaneous reporting



LAERTES Evidence Map

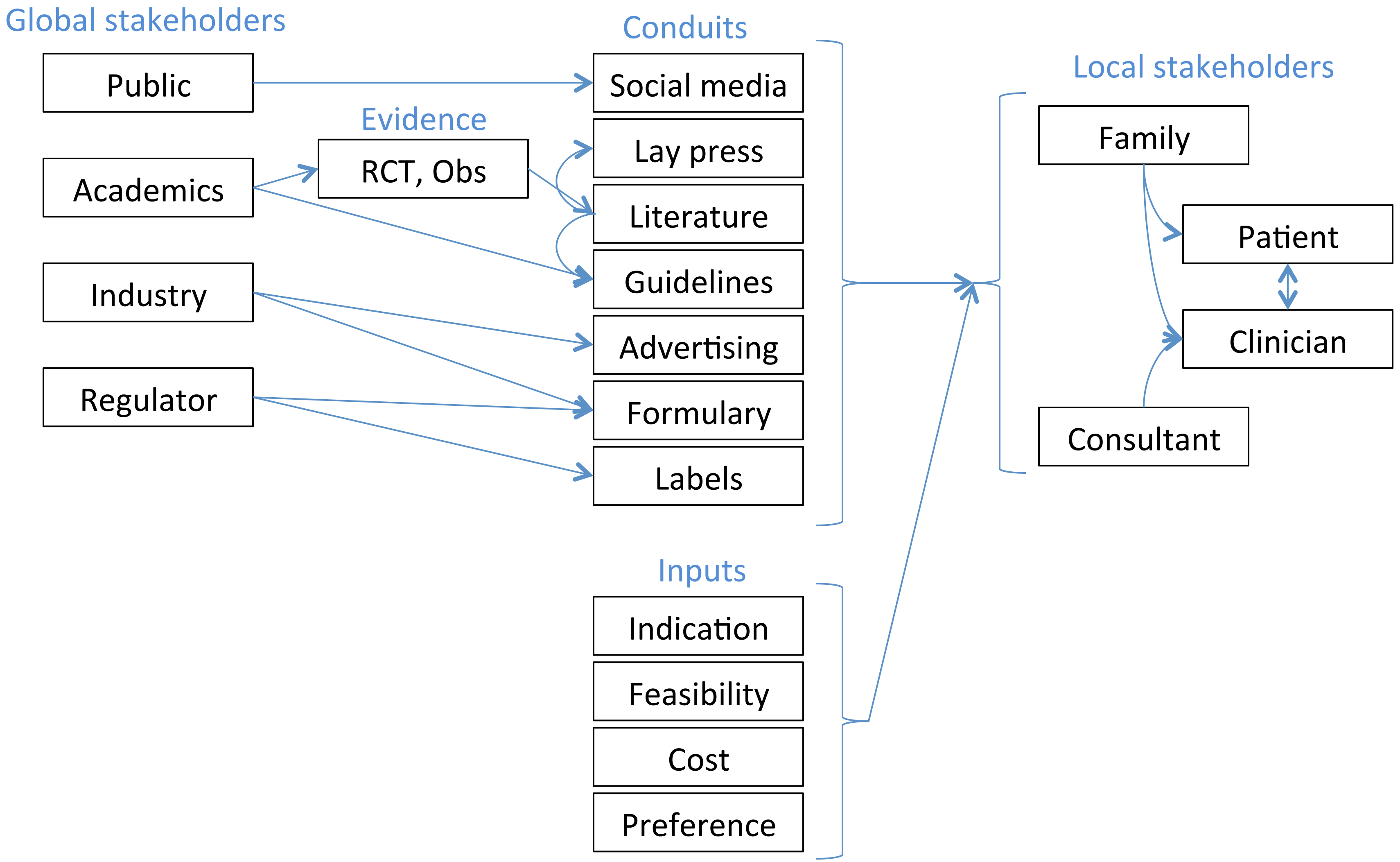
OHDSI in Action

- Generate evidence
  - Randomized trial is the gold standard
  - Observational research is supporting
    - Can it become a partnership?

Characterization

- Today we carry out RCTs without clear knowledge of actual practice
- There will be no RCTs without an observational precursor
  - It will be required to characterize a population using large-scale observational data before designing an RCT
  - Disease burden
  - Actual treatment practice
  - Time on therapy
  - Course and complication rate
  - Done now somewhat through literature and pilot studies

Treatment Pathways



Network process

1. Join the collaborative
2. Propose a study to the open collaborative
3. Write protocol
   - http://www.ohdsi.org/web/wiki/doku.php?id=research:studies
4. Code it, run it locally, debug it (minimize others' work)
5. Publish it: https://github.com/ohdsi
6. Each node voluntarily executes on their CDM
7. Centrally share results
8. Collaboratively explore results and jointly publish findings

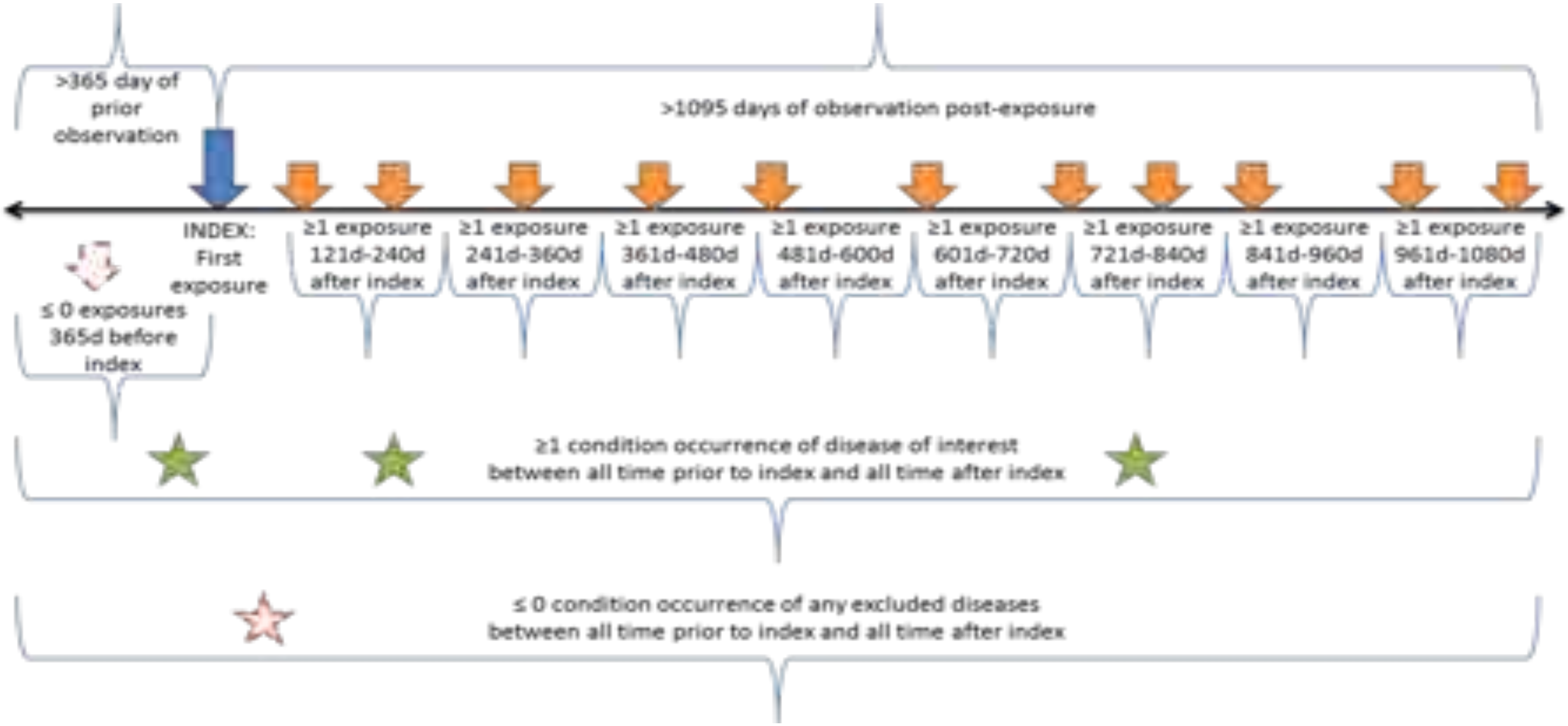
OHDSI in action:
Chronic disease treatment pathways

• Conceived at AMIA	15Nov2014
• Protocol written, code written and tested at 2 sites	30Nov2014
• Analysis submitted to OHDSI network	2Dec2014
• Results submitted for 7 databases	5Dec2014

OHDSI participating data partners

Abbre-viation	Name	Description	Population, millions
AUSOM	Ajou University School of Medicine	South Korea; inpatient hospital EHR	2
CCAE	MarketScan Commercial Claims and Encounters	US private-payer claims	119
CPRD	UK Clinical Practice Research Datalink	UK; EHR from general practice	11
CUMC	Columbia University Medical Center	US; inpatient EHR	4
GE	GE Centricity	US; outpatient EHR	33
INPC	Regenstrief Institute, Indiana Network for Patient Care	US; integrated health exchange	15
JMDC	Japan Medical Data Center	Japan; private-payer claims	3
MDCD	MarketScan Medicaid Multi-State	US; public-payer claims	17
MDCR	MarketScan Medicare Supplemental and Coordination of Benefits	US; private and public-payer claims	9
OPTUM	Optum ClinFormatics	US; private-payer claims	40
STRIDE	Stanford Translational Research Integrated Database Environment	US; inpatient EHR	2
HKU	Hong Kong University	Hong Kong; EHR	1

Treatment pathway event flow



>365 day of prior observation
>1095 days of observation post-exposure
INDEX: First exposure
≥1 exposure 121d-240d after index
≥1 exposure 241d-360d after index
≥1 exposure 361d-480d after index
≥1 exposure 481d-600d after index
≥1 exposure 601d-720d after index
≥1 exposure 721d-840d after index
≥1 exposure 841d-960d after index
≥1 exposure 961d-1080d after index
≤ 0 exposures 365d before index
≥1 condition occurrence of disease of interest between all time prior to index and all time after index
≤ 0 condition occurrence of any excluded diseases between all time prior to index and all time after index

Proceedings of the National Academy of Sciences, 2016

Characterizing treatment pathways at scale using the OHDSI network

George Hripcsak[a,b,c,1], Patrick B. Ryan[c,d], Jon D. Duke[c,e], Nigam H. Shah[c,f], Rae Woong Park[c,g], Vojtech Huser[c,h], Marc A. Suchard[c,i,j,k], Martijn J. Schuemie[c,d], Frank J. DeFalco[c,d], Adler Perotte[a,c], Juan M. Banda[c,f], Christian G. Reich[c,l], Lisa M. Schilling[c,m], Michael E. Matheny[c,n,o], Daniella Meeker[c,p,q], Nicole Pratt[c,r], and David Madigan[c,s]

[a]Department of Biomedical Informatics, Columbia University Medical Center, New York, NY 10032; [b]Medical Informatics Services, NewYork-Presbyterian Hospital, New York, NY 10032; [c]Observational Health Data Sciences and Informatics, New York, NY 10032; [d]Epidemiology Analytics, Janssen Research and Development, Titusville, NJ 08560; [e]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN 46205; [f]Center for Biomedical Informatics Research, Stanford University, CA 94305; [g]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea, 443-380; [h]Lister Hill National Center for Biomedical Communications (National Library of Medicine), National Institutes of Health, Bethesda, MD 20894; [i]Department of Biomathematics, University of California, Los Angeles, CA 90095; [j]Department of Biostatistics, University of California, Los Angeles, CA 90095; [k]Department of Human Genetics, University of California, Los Angeles, CA 90095; [l]Real World Evidence Solutions, IMS Health, Burlington, MA 01803; [m]Department of Medicine, University of Colorado School of Medicine, Aurora, CO 80045; [n]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37212; [o]Geriatric Research, Education and Clinical Center, VA Tennessee Valley Healthcare System, Nashville, TN 37212; [p]Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90089; [q]Department of Pediatrics, University of Southern California, Los Angeles, CA 90089; [r]Division of Health Sciences, University of South Australia, Adelaide, SA, Australia 5001; and [s]Department of Statistics, Columbia University, New York, NY 10027

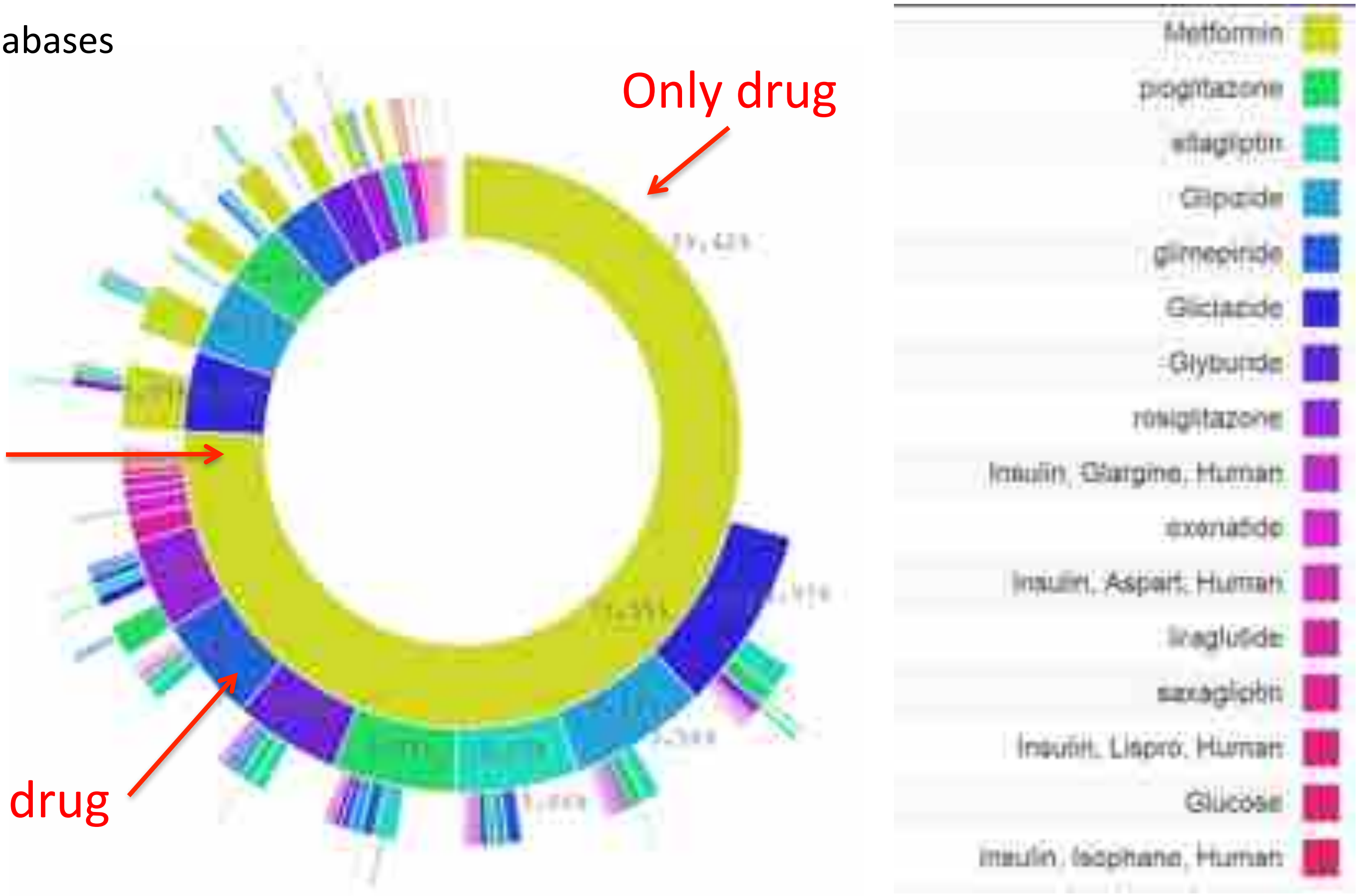
Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved April 5, 2016 (received for review June 14, 2015)

Observational research promises to complement experimental research by providing large, diverse populations that would be infeasible for an experiment. Observational research can test its own clinical hypotheses, and observational studies also can contribute to the design of experiments and inform the generalizability of experimental research. Understanding the diversity of populations

Without sufficiently broad databases available in the first stage, randomized trials are designed without explicit knowledge of actual disease status and treatment practice. Literature reviews are restricted to the population choices of previous investigations, and pilot studies usually are limited in scope. By exploiting the ClinicalTrials.gov national trial registry (9) and electronic health

Treatment pathways for diabetes

T2DM : All databases

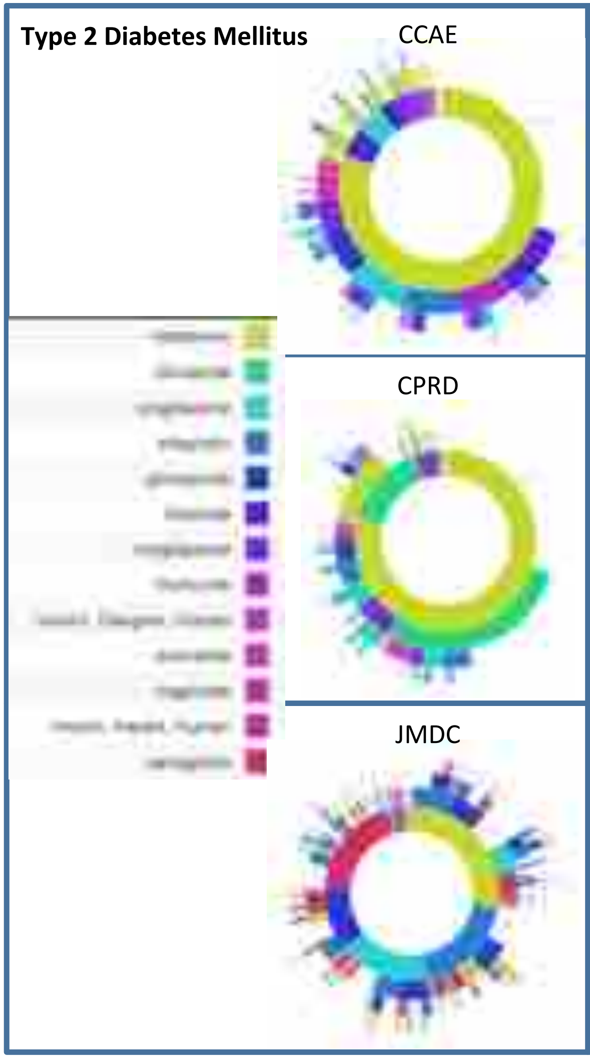


Only drug

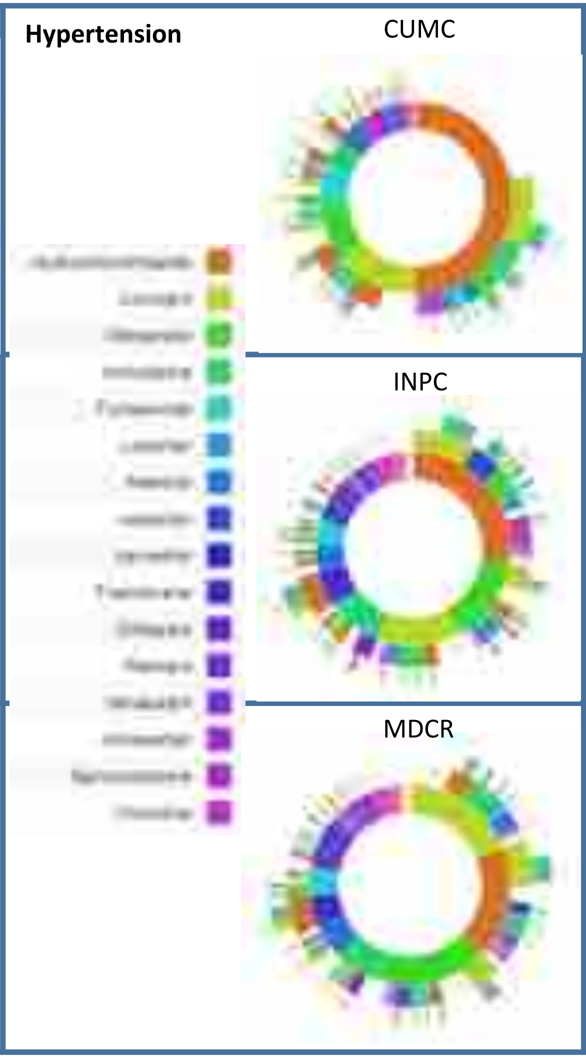
First drug

Second drug

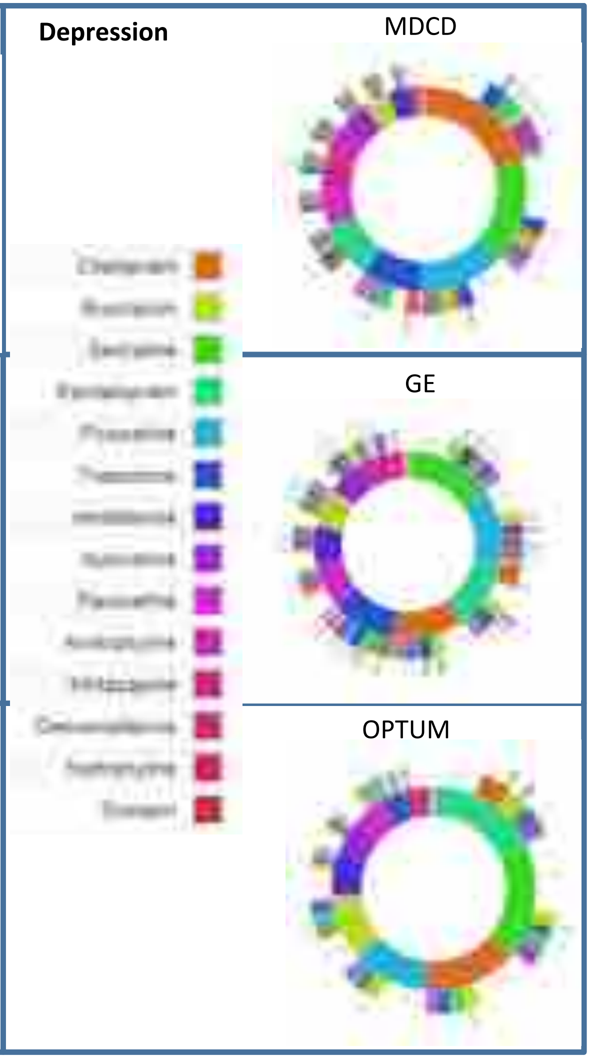
Population-level heterogeneity across systems, and patient-level heterogeneity within systems



Type 2 Diabetes Mellitus



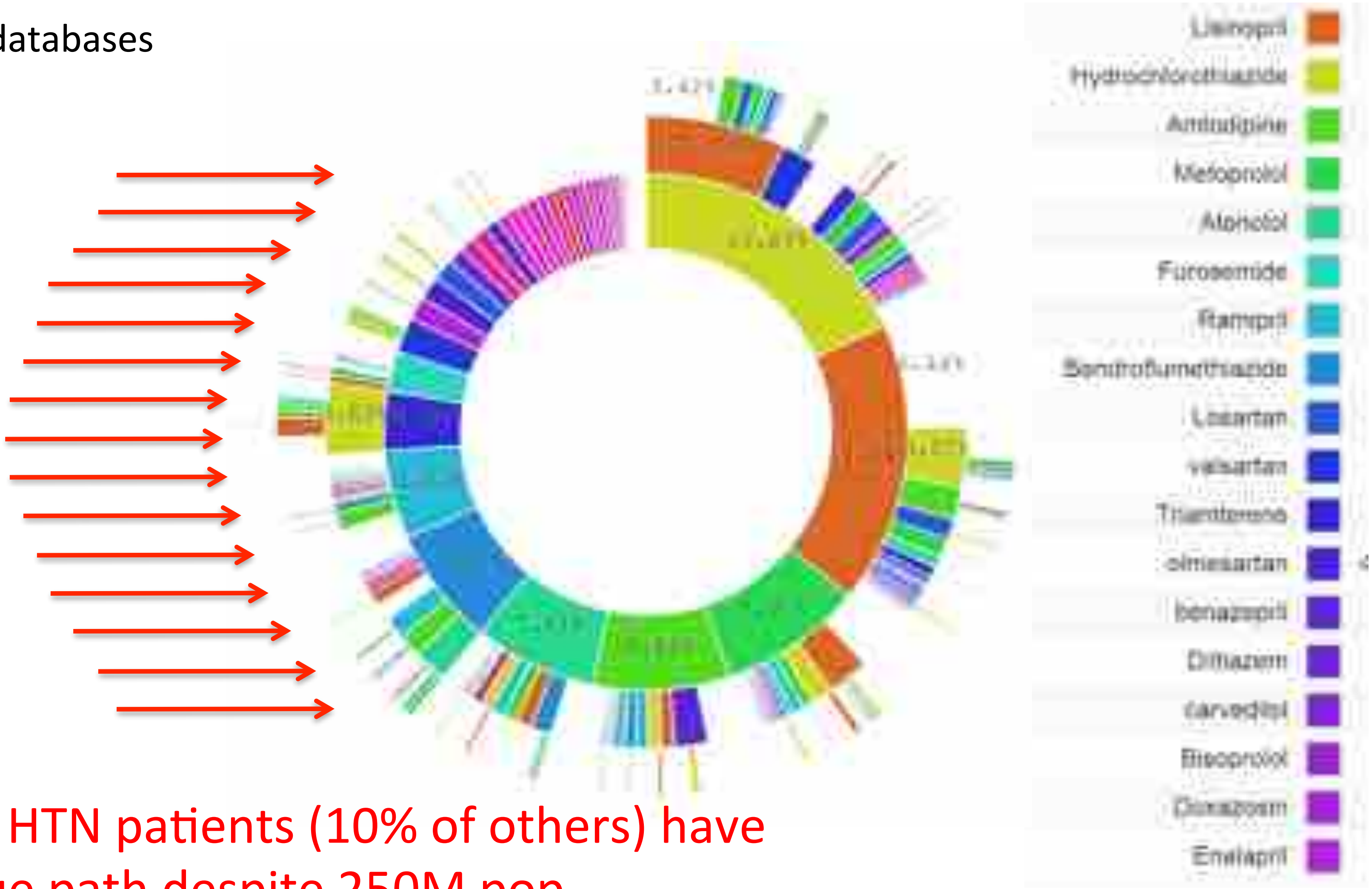
Hypertension



Depression

Patient-level heterogeneity

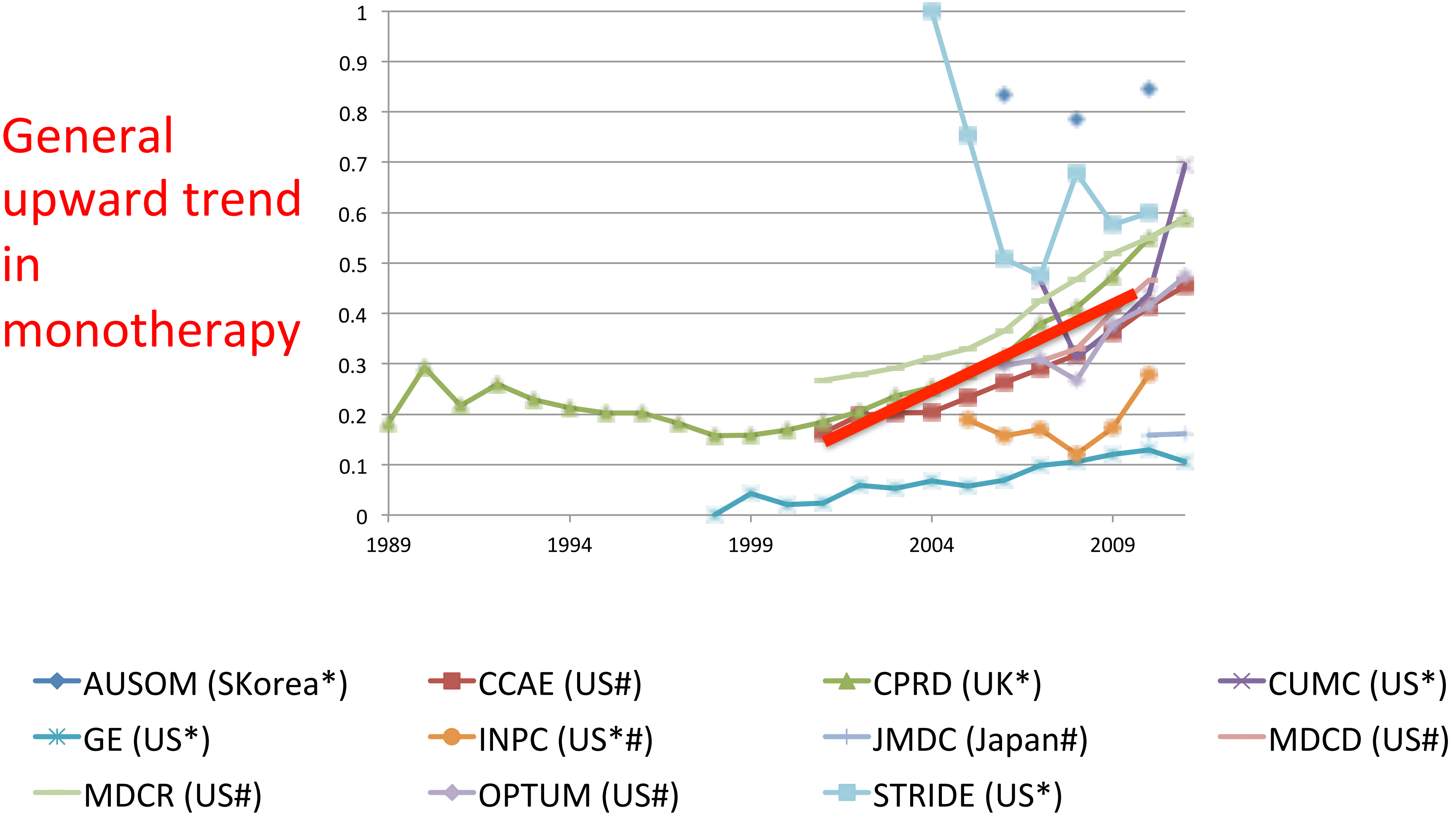
HTN: All databases



25% of HTN patients (10% of others) have a unique path despite 250M pop

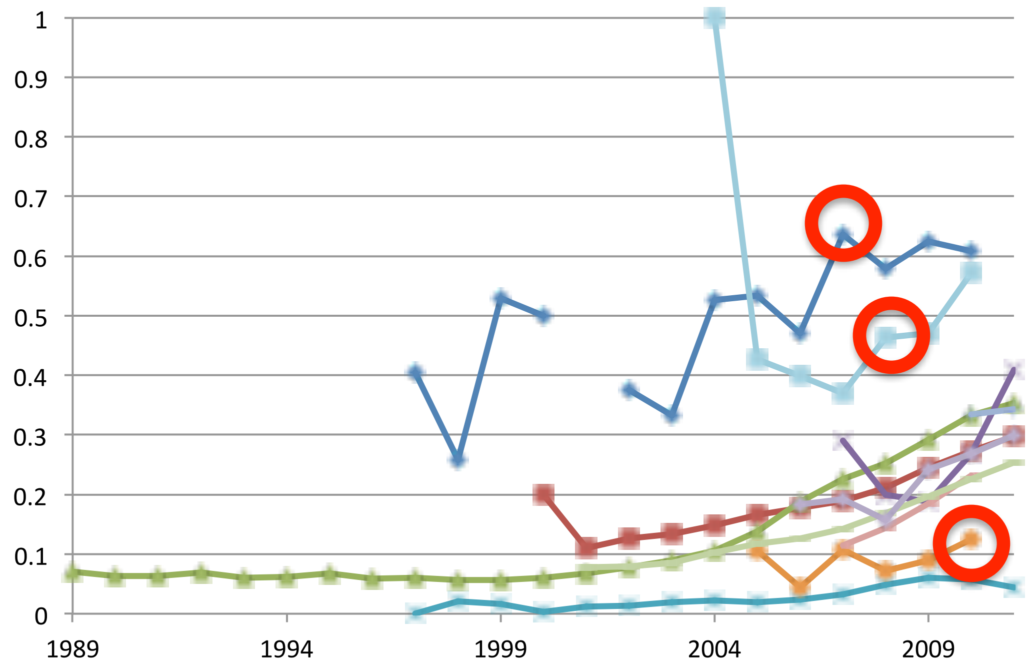
Monotherapy – diabetes

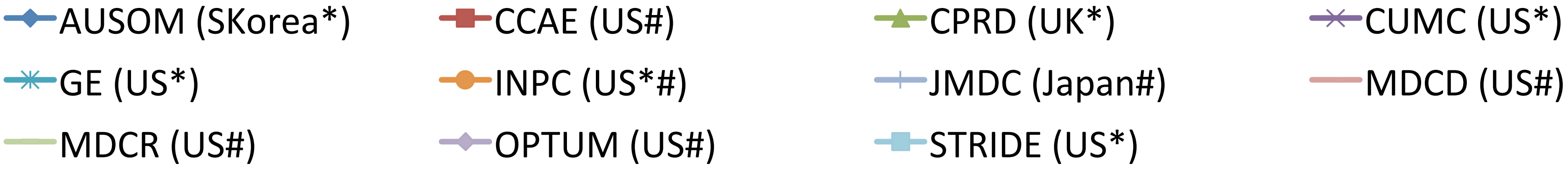
General upward trend in monotherapy



Monotherapy – HTN

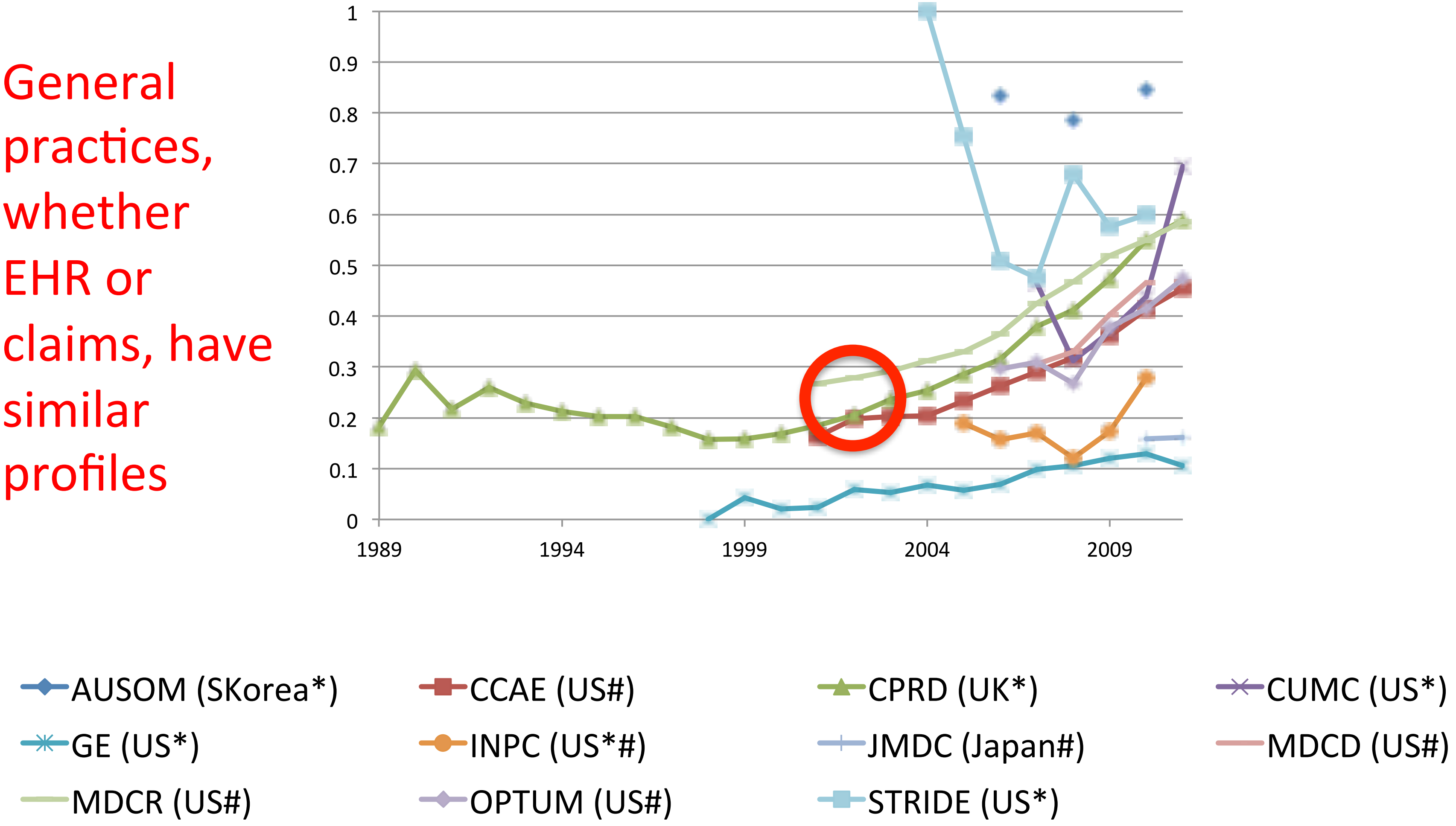
Academic medical centers differ from general practices





Monotherapy – diabetes

General practices, whether EHR or claims, have similar profiles



Conclusions: Network research

- It is feasible to encode the world population in a single data model
  - Over 600,000,000 records by voluntary effort (682,000,000)
- Generating evidence is feasible
- Stakeholders willing to share results
- Able to accommodate vast differences in privacy and research regulation