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

http://ohdsi.org
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

All health outcomes of interest

All drugs
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 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

Population-level estimation for comparative effectiveness research:

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

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

Source data warehouse, with identifiable patient-level data

ETL

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

Standardized large-scale analytics

Analysis results

OHDSI Coordinating Center

Data network support

Analytics development and testing

Research and education

OHDSI.org

Summary statistics results repository

OHDSI Data Partners

Biological gradient

Specificity

Analogy

Comparative effectiveness

Predictive modeling

Consistency

Temporal

Strength

Plausibility

Experiment

Coherence
Deep information model
OMOP CDM v5.0.1

Person

Standardized clinical data

- 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
- Payer_plan_period
- Cost

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

Standardized derived elements
Extensive vocabularies
Preparing your data for analysis

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
### Data Quality Messages

- **Message Type:** ERRGR
  - **Message:** 101 - Number of persons by age, with age at first observation period; should not have age < 0, (n=9480)

- **Message Type:** ERRGR
  - **Message:** 103 - Distribution of age at first observation period (count = 1); min value should not be negative

- **Message Type:** ERRGR
  - **Message:** 114 - Number of persons with observation period before year of birth; count (n=851) should not be > 0

- **Message Type:** ERRGR
  - **Message:** 206 - Distribution of age by visit_concept_id (count = 7); min value should not be negative

- **Message Type:** ERRGR
  - **Message:** 301 - Number of providers by specialty_concept_id; 224 concepts in data are not in correct vocabulary (Specialty)

- **Message Type:** ERRGR
  - **Message:** 400 - Number of persons with at least one condition occurrence by condition_concept_id; 113 concepts in data are not in correct vocabulary (ICD-10)

- **Message Type:** ERRGR
  - **Message:** 406 - Distribution of age by condition_concept_id; (count = 763); min value should not be negative
ATLAS to build, visualize, and analyze cohorts
Characterize the cohorts of interest
LAERTES: Knowledge base of what we know: literature, labeling, spontaneous reporting
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

Global stakeholders

- Public
- Academics
- Industry
- Regulator

Evidence

- RCT, Obs

Conduits

- Social media
- Lay press
- Literature
- Guidelines
- Advertising
- Formulary
- Labels

Local stakeholders

- Family
- Patient
- Clinician
- Consultant

Inputs

- Indication
- Feasibility
- Cost
- Preference

Local stakeholders

- Family
- Patient
- Clinician
- Consultant
Network process

1. Join the collaborative
2. Propose a study to the open collaborative
3. Write protocol
4. Code it, run it locally, debug it (minimize others’ work)
5. Publish it: [https://github.com/ohdsi](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

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

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

≥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
Characterizing treatment pathways at scale using the OHDSI network

George Hripcsak\textsuperscript{1,1}, Patrick B. Ryan\textsuperscript{1,3}, Jon O. Duke\textsuperscript{1,4}, Nigam H. Shah\textsuperscript{1,5}, Iae Woong Park\textsuperscript{1,6}, Vojtech Huter\textsuperscript{1,7}, Marc A. Suchard\textsuperscript{1,1,8}, Martijn J. Schuemie\textsuperscript{1,9}, Frank J. DeFalco\textsuperscript{1,10}, Adler Perotte\textsuperscript{1,11}, Juan M. Banda\textsuperscript{1,11}, Christian G. Reich\textsuperscript{1,11}, Lisa M. Schilling\textsuperscript{1,12}, Michael E. Matheny\textsuperscript{1,13}, Daniella Meeker\textsuperscript{1,14}, Nicole Pratt\textsuperscript{1,15}, and David Maltiga\textsuperscript{1,16}

\textsuperscript{1}Department of Biomedical Informatics, Columbia University Medical Center, New York, NY 10032; \textsuperscript{2}Medical Informatics Service, New York Presbyterian Hospital, New York, NY 10022; \textsuperscript{3}Observational Health Data Sciences and Informatics, New York, NY 10022; \textsuperscript{4}Epidemiology, Jersey Research and Development, Edgewater, NJ 07020; \textsuperscript{5}Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN 46202; \textsuperscript{6}Center for Biomedical Informatics Research, Stanford University, CA 94305; \textsuperscript{7}Department of Biomedical Informatics, AUS University School of Medicine, Sousse, South Korea 443-160; \textsuperscript{8}Lakerill National Center for Biomedical Communications (National Library of Medicine, National Institute of Health, Bethesda, MD 20894); Department of Biostatistics, University of California, Los Angeles, CA 90095; \textsuperscript{9}Department of Biostatistics, University of California, Los Angeles, CA 90095; \textsuperscript{10}Real World Evidence Solutions, IMS Health, Burlington, MA 01803; \textsuperscript{11}Department of Medicine, University of Colorado School of Medicine, Aurora, CO 80045; \textsuperscript{12}Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232; \textsuperscript{13}Scripps Research, Education and Clinical Center, VA Tennessee Valley Healthcare System, Nashville, TN 37212; \textsuperscript{14}Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90033; \textsuperscript{15}Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90033; \textsuperscript{16}Division of Health Sciences, University of South Australia, Adelaide, SA, Australia 5000; and Department of Nephrology, Columbia University, New York, NY 10032.

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 other 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

First drug

Second drug

Only drug
Population-level heterogeneity across systems, and patient-level heterogeneity within systems
Patient-level heterogeneity

HTN: All databases

25% of HTN patients (10% of others) have a unique path despite 250M pop
General upward trend in monotherapy
Monotherapy – HTN

Academic medical centers differ from general practices
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