Observational Health Data Sciences and Informatics, Interoperability, and Research

George Hripcsak, MD, MS
Columbia University Irving Medical Center
NewYork-Presbyterian Hospital
Observational Health Data Sciences and Informatics
Observational Health Data Sciences and Informatics (OHDSI, as “Odyssey”)

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

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

http://ohdsi.org
OHDSI’s global research community

- >300 collaborators from 30 different countries
- Experts in informatics, statistics, epidemiology, clinical sciences
- Active participation from academia, government, industry, providers
- Records on about 600 million unique patients in >100 databases

http://ohdsi.org/who-we-are/collaborators/
Evidence OHDSI seeks to generate from observational data

• Clinical characterization - tally
  – Natural history: Who has diabetes, and who takes metformin?
  – Quality improvement: What proportion of patients with diabetes experience complications?

• Population-level estimation - cause
  – Safety surveillance: Does metformin cause lactic acidosis?
  – Comparative effectiveness: Does metformin cause lactic acidosis more than glyburide?

• Patient-level prediction - predict
  – 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?
Open Science

Data + Analytics + Domain expertise

Open source software

Enable users to do something

Generate evidence

Standardized, transparent workflows

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

OHDSI Coordinating Center

Data network support

Analytics development and testing

Research and education

Summary statistics results repository

OHDSI.org

OHDSI Data Partners

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 Data Partners

OHDSI Coordinating Center

Data network support

Analytics development and testing

Research and education

Summary statistics results repository

OHDSI.org
What's the adherence to my drug in the data assets I own?

Current Approach: “One Study – One Script”

Analytical method: Adherence to Drug

Application to data

Current solution:

Custom script for each study

- Not scalable
- Expensive
- Slow
- Prohibitive to non-expert routine use
Solution: Standardized Data and Analytics

1. ATLAS, Remote Studies
   - Standard Cohorts
   - Standardized Analytics

2. OMOP CDM
   - Standardized Format
   - Standardized Coding
Common Data Model

• OMOP
  – Observational Medical Outcomes Partnership
  – (Origin of OHDSI; kept the data model name)

• Components
  – Schema – tables where you put data
  – Vocabulary – what codes go in the table
  – Conventions – how to store data

• Open committee structure to govern it
  – Contracted vocabulary maintenance
Deep information model
OMOP CDM

Standardized clinical data
- Person
  - Observation_period
  - Visit_occurrence
    - Visit_detail
  - Condition_occurrence
  - Drug_exposure
  - Procedure_occurrence
  - Device_exposure
  - Measurement
  - Note
    - Note_NLP
  - Survey_conduct
  - Observation
    - Specimen
    - Fact_relationship

Standardized derived elements
- Condition_era
- Drug_era
- Dose_era

Results Schema
- Cohort
  - Cohort_definition
- CDM_source

Standardized health system data
- Location
- Location_history
- Care_site
- Provider

Standardized metadata
- CDM_source
- Metadata

Standardized vocabularies
- Concept
- Vocabulary
- Domain
- Concept_class
- Concept_relationship
- Relationship
- Concept_synonym
- Concept_ancestor
- Source_to_concept_map
- Drug_strength

Standardized health economics
- Cost
- Payer_plan_period
Extensive vocabularies

Breakdown of OHDSI concepts by domain, standard class, and vocabulary
OHDSI’s standardized vocabularies

• 153 Vocabularies across 41 domains
  – MU3 standards: SNOMED, RxNorm, LOINC
  – Disparate sources: ICD9CM, ICD10(CM), Read, NDC, Gemscript, CPT4, HCPCS...

• >9 million concepts
  – >3.3 million standard concepts
  – >5.1 million source codes
  – >629,000 classification concepts

• >55 million concept relationships
• >84 million ancestral relationships

Publicly available for download at: http://athena.ohdsi.org/
Standard vocabularies

• Bring the world’s data to a core set
• RxNorm
  – RxNorm Extensions - cover non-US
• LOINC
• SNOMED CT
  – How to address non-US data
• In many areas, stuck adopting several
  – Procedures, ...
Standardized conventions

Shared Conventions developed by the THEMIS Workgroup

203

Standardized clinical data

- Person
  - Observation_period
  - Visit_occurrence
    - Visit_detail
  - Condition_occurrence
  - Drug_exposure
  - Procedure_occurrence
  - Device_exposure
  - Measurement
  - Note
    - Note_NLP
  - Survey_conduct
  - Observation
  - Specimen
  - Fact_relationship

Standardized health system data

- Location
  - Location_history
- Care_site
- Provider

Standardized derived elements

- Condition_era
- Drug_era
- Dose_era

Results Schema

- Cohort
- Cohort_definition

Standardized health economics

- Cost
- Payer_plan_period

Standardized metadata

- CDM_source
- Metadata
Preparing your data

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](http://github.com/OHDSI)
Data Quality Dashboard

IBM® MARKETSCAN® MULTI-STATE MEDICAID DATABASE

DataQualityDashboard Version: 1.0.0
Results generated at 2020-08-24 15:44:34 in 3 hours

<table>
<thead>
<tr>
<th>STATUS</th>
<th>TABLE</th>
<th>CATEGORY</th>
<th>SUBCATEGORY</th>
<th>LEVEL</th>
<th>NOTES</th>
<th>DESCRIPTION</th>
<th>% RECORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAIL</td>
<td>PAYER_PLAN_PERIOD</td>
<td>Conformance</td>
<td>Relational</td>
<td>FIELD</td>
<td>None</td>
<td>The number and percent of records that have a value in the payer_plan_period_id field in the PAYER_PLAN_PERIOD table that does not exist in the PERSON table. (Threshold=0%).</td>
<td>100.00%</td>
</tr>
<tr>
<td>FAIL</td>
<td>PROVIDER</td>
<td>Conformance</td>
<td>None</td>
<td>FIELD</td>
<td>None</td>
<td>The number and percent of records that do not have a standard, valid concept in the gender_concept_id field in the PROVIDER table. (Threshold=0%).</td>
<td>100.00%</td>
</tr>
<tr>
<td>PASS</td>
<td>PERSON</td>
<td>Completeness</td>
<td>None</td>
<td>FIELD</td>
<td>None</td>
<td>The number and percent of records with a NULL value in the birth_date_time of the PERSON. (Threshold=100%).</td>
<td>100.00%</td>
</tr>
<tr>
<td>PASS</td>
<td>PERSON</td>
<td>Completeness</td>
<td>None</td>
<td>FIELD</td>
<td>None</td>
<td>The number and percent of records with a NULL value in the provider_id of the PERSON. (Threshold=100%).</td>
<td>100.00%</td>
</tr>
<tr>
<td>PASS</td>
<td>PERSON</td>
<td>Completeness</td>
<td>None</td>
<td>FIELD</td>
<td>None</td>
<td>The number and percent of records with a NULL value in the care_site_id of the PERSON. (Threshold=100%).</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Showing 1 to 5 of 3,124 entries
ATLAS: Ontology support

- What terms do I need to create a cohort
- Tied to the database: what terms are used
  - Especially important for someone else’s database
ATLAS: Cohort building

- Optimized for observational research
  - Time series: who and when (vs classification)
  - Observation period, event timing
  - Assume a complex definition – Linearized AND-OR group
ATLAS: Visualization

- Tables
- Graphs
ATLAS: Analysis (observational)

- Approach: log regression, Poisson regression, survival
- Confounder: regularized-regression propensity score
- Residual confounding: calibration
- Diagnostics
OHDSI in Action
Clinical Practice Guideline: Executive Summary


A Report of the American College of Association Task Force on Clinical Practice Guidelines

WRITING COMMITTEE
Paul K. Whelton, MB, MD, MSc, FAHA, Chair; R
Wilbert S. Aronow, MD, FACC, FAHA; Donald E. Casey, Jr, M
Cheryl Dennison Himmelfarb, RN, ANP, PhD, FAHA; So
Samuel Gidding, MD, FAHA; Kenneth A. Jamerson
Eric J. MacLaughlin, PharmD; Paul Muntner, PhD, FAHA; Sidd
Sidney C. Smith, Jr, MD, MACC, FAHA; Crystal C. S
Sandu J. Taler, MD, FAHA; Randall J. Thomas, MD, MS, FACC
Jeff D. Williamson, MD, MHS; Jackson T.

ACC/AHA TASK FORCE
Glenn N. Levine, MD, FACC, FAHA, Chair; Patrick T.
Jonathan L. Halperin, MD, FACC, FAHA, Immediate Past Chi
Joshua A. Beckman, MD, MS, FAHA; Kim K. Ritchie, MS, PharmD
Ralph G. Brindis, MD, MPH, MACC; Joaquín E. Cigars
Anita Deswal, MD, MPH, FACC, FAHA; Lee A. Fleisher, M
Samuel Gidding, MD, FAHA; Zachary D. Goldberger, MD, MS
John Ikonomidis, MD, PhD, FAHA; José A. Jodlar, MD, F
Susan J. Pressler, PhD, RN, FAHA; Barbara Riegel, PhD,

Head-to-head HTN drug comparisons

- Trials: 40
- \( N = 102 - [1148] - 33 \)K

- Comparisons: 10,278
- \( N = 3502 - [212K] - 1.9 \)M

OHDSI “LEGEND” Hypertension Study
Filling in the evidence gaps
Hypertension, cardiac research

THE LANCET

Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis

Marie A Suchard, Anjali J Schmeker, Helen Al Kronish, Seng Chan You, Alvin Chen, Nanci Pratt, Christian G Reisch, Jin-Ok Lee, David Madigan, George Hipscoak, Patrick B Ryan

Summary
Background Uncertainty remains about the optimal monotherapy for hypertension, with current guidelines recommending any primary agent among the first-line drug classes thiazide or thiazide-like diuretics, angiotensin-converting

Hypertension

BETA-BLOCKER THERAPY

Comprehensive Comparative Effectiveness and Safety of First-Line β-Blocker Monotherapy in Hypertensive Patients
A Large-Scale Multicenter Observational Study

Seng Chan You, Helen M Kronish, Marc A Suchard, Martin J Schmaeker, George Hipsck, Paul J Chen, Steven Shef, Jean-Duc Le, Nanci Pratt, Christian G Reisch, Patrick B Ryan, Chan Hwang Park, Sung H Park

ABSTRACT: Evidence for the effectiveness and safety of the third-generation β-blockers other than atenolol in hypertension remains sparse. We assessed the effectiveness and safety of β-blockers as first-line treatment for hypertension using 3

JAMA Internal Medicine

Comparison of Cardiovascular and Safety Outcomes of Chlorthalidone vs Hydrochlorothiazide to Treat Hypertension
George Hipsck, MD, MS, Marc A Suchard, MD, PhD, Steven Shef, MD, Burton Chen, MD, Seng Chan You, MD, Nicole Pratt, PhD, David Madigan, PhD, Helen M Kronish, MD, IMA
Patrick B Ryan, PhD, Martin J Schmaeker, PhD

IMPORTANCE: Chlorthalidone is currently recommended as the preferred thiazide diuretic to treat hypertension, but no trials have directly compared risks and benefits.

OBJECTIVE: To compare the effectiveness and safety of chlorthalidone and hydrochlorothiazide as first-line therapies for hypertension in real-world practice.

JAMA

Association of Ticagrelor vs Clopidogrel With Net Adverse Clinical Events in Patients With Acute Coronary Syndrome Undergoing Percutaneous Coronary Intervention
Seng Chan You, MD, MS, Yoonsook Ri, PhD, Behnam Bikelid, MD, MS, Jiwoo Kim, MS, Anestasios Sapos, MSC, James Weinzer, MSC, A.J. Londo, MPH, Jeeyong Che, BS, Jihyung Park, BS, Martin J Schmaeker, PhD, Marc A Suchard, MD, PhD, David Madigan, PhD, George Hipsck, MD, MS, Atsefr Gabor, MD, MS, Christian G Reisch, MD, Patrick B Ryan, PhD, Bao-Huang Park, MD, PhD, Helen M Kronish, MD, IMA

IMPORTANCE: Current guidelines recommend ticagrelor as the preferred P2Y12 inhibitor
COVID-19 Research

• 4.5 million COVID-19 patient records
  – Academic centers (CDW)
  – EHR and claims data aggregators
  – Government databases

• 41 studies carried out in past year
Evidence was needed around the use of hydroxychloroquine (HCQ) alone and in combination with azithromycin (AZ). We examined the use of these drugs in rheumatoid arthritis (RA) patients.

Findings:
- In history use in RA population, HCQ alone is generally safe but in combination with AZ it shows a doubling of risk of 30-day cardiovascular mortality.
Patients with cardiovascular diseases and hypertension treated with angiotensin converting enzyme inhibitors (ACEs) angiotensin-II receptor blockers (ARBs) may influence susceptibility to COVID-19 and worsen its severity.
COVID-19 Vaccine Safety Methods Research

- **AstraZeneca vaccine**
  - March 11-15, 2021 – 13 European countries suspend use for fears of blood clots
    - Denmark, Norway, Iceland, Bulgaria, Ireland, Netherlands, Spain, Germany, Italy, France, ...
  - As of March 16, 2021 – of 20 million persons vaccinated in Europe several deaths
    - 469 thromboembolic events reported after vaccination
    - 7 cases disseminated intravascular coagulation (DIC)
    - 18 cases cerebral venous sinus thrombosis (CVST)
  - March 18, 2021 – EMA determines benefits outweigh the risks
    - Thromboembolic events “lower than that expected in the general population”
    - DIC and CVST above baseline but very rare
    - “The number of reported events exceeds those expected, and causality although not confirmed, cannot therefore be excluded. However, given the rarity of the events, and the difficulty of establishing baseline incidence since COVID-19 itself is resulting in hospitalisations with thromboembolic complications, the strength of any association is uncertain.”

- Partnered with FDA Center for Biologics Evaluation and Research (CBER)
  - Vaccine safety methods research, network and local studies
OHDSI OMOP Common Data Model

**FDA CBER BEST Program**

**INTRODUCTION**
The U.S. Food and Drug Administration (FDA) is collaborating with the Office of Biologics and Medical Products (OBI) to develop a comprehensive database to support the best practices in the development and evaluation of vaccine candidates.

**OBJECTIVE**
The primary objective of this database is to improve vaccine development by providing a comprehensive and structured database of vaccine candidates.

**METHODS**
The database will collect and analyze data from various sources, including clinical trials, animal models, and epidemiological studies.

**All of Us Research Program**

The All of Us Research Program is a historic effort to gather data from one million or more persons living in the United States to accelerate research and improve health. By taking into account individual differences in lifestyle, environment, and biology, researchers will uncover paths toward delivering precision medicine.

**N3C National COVID Cohort**

The N3C is a national cohort of over 4 million participants with health data linked to National Institutes of Health (NIH)-funded clinical research networks.

**eMERGE Network**

**Collaborations:** PCORnet, S4S, ...
HL7 – OHDSI Partnership

• Announced March 1, 2021
  – FHIR and OMOP
  – “The organizations will align their standards to capture data in a clearly defined way into a single common data model. This will allow clinicians as well as researchers to pull data from multiple sources and compile it in the same structure without degradation of the information.”

• Early in creation of joint working groups and scope
  – Starting with existing mapping work
    • Georgia Tech, EHDEN, MIRACUM, Leiden University, Denmark CSS

• Range
  – Mapping, shared knowledge engineering, common standard

• Welcome the feedback
Research as “secondary use”

• Research is not an afterthought
  – It drives health care and saves millions of lives
• Jonas Salk invented the polio vaccine
  – No one remembers his billing records
• Without research we would be billing for leeches
• Perhaps (after patient care) research should become paramount and billing should make due
Mission

• You can store data or you can generate evidence, but you cannot do both
Coupling

- Therefore, need a tightly coupled enterprise, evidence generation & standards development
  - OHDSI governance structure
  - Need analytic standards as well as data
  - Standard must be connected to community
  - Open source
It’s not magic

• Someone has to pay the data-quality price
  – In All of Us RP, found conversion is best close to data generation
  – FHIR dump of all data to a distance central warehouse will be tough
US-specific standards

- US is 5% of the population
  - Continued focus on US standards hurts US citizens

- Modern causal inference is data hungry
  - Cannot do most research on just the US population