Data Availability and Benchmarking in Observational Health Data Sciences and Informatics (OHDSI)

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

International Telecommunication Union & World Health Organization
Artificial Intelligence for Health
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

- >200 collaborators from 25 different countries
- Experts in informatics, statistics, epidemiology, clinical sciences
- Active participation from academia, government, industry, providers
- Currently records on about 500 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

Database summary

Cohort definition

Cohort summary

Compare cohorts

Exposure-outcome summary

Effect estimation & calibration

Compare databases

Standardized, transparent workflows
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
Extensive vocabularies

Breakdown of OHDSI concepts by domain, standard class, and vocabulary
Standardized conventions developed by the THEMIS Workgroup
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
<table>
<thead>
<tr>
<th>Message Type</th>
<th>Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERROR</td>
<td>101 - Number of persons by age, with age at first observation period; should not have age &lt; 0, (n=848)</td>
</tr>
<tr>
<td>ERROR</td>
<td>103 - Distribution of age at first observation period (count = 1); min value should not be negative</td>
</tr>
<tr>
<td>ERROR</td>
<td>114 - Number of persons with observation period before year-of-birth; count (n=851) should not be &gt; 0</td>
</tr>
<tr>
<td>ERROR</td>
<td>206 - Distribution of age by visit_concept_id (count = 7); min value should not be negative</td>
</tr>
<tr>
<td>ERROR</td>
<td>301 - Number of providers by specialty concept_id; 224 concepts in data are not in correct vocabulary (Specialty)</td>
</tr>
<tr>
<td>ERROR</td>
<td>400 - Number of persons with at least one condition occurrence, by condition_concept_id; 115 concepts in data are not in correct vocabulary (SNOMED)</td>
</tr>
<tr>
<td>ERROR</td>
<td>406 - Distribution of age by condition_concept_id (count = 753); min value should not be negative</td>
</tr>
</tbody>
</table>
ATLAS to build, visualize, and analyze cohorts
Characterize the cohorts of interest
OHDSI in Action

• Characterization
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
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>
Characterizing treatment pathways at scale using the OHDSI network

George Hripcsak e, f, g, h, Patrick B. Ryan e, f, g, h, Jon D. Duke e, f, Nigam H. Shah e, f, Rae Woong Park e, f, g, Vojtech Huser e, f, Marc A. Suchard e, f, g, h, Martijn J. Schuemie e, f, Frank J. DeFalco e, f, Adler Perotte e, g, h, Juan M. Banda e, f, Christian G. Reich e, f, Lisa M. Schilling e, f, Michael E. Matheny e, f, g, h, Daniella Meeker e, f, g, h, Nicole Pratt e, f, and David Madigan e, f, g

Department of Biomedical Informatics, Columbia University Medical Center, New York, NY 10032; Medical Informatics Services, NewYork-Presbyterian Hospital, New York, NY 10032; Observational Health Data Sciences and Informatics, New York, NY 10032; Epidemiology Analytics, Janssen Research and Development, Titusville, NJ 08560; Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN 46205; Center for Biomedical Informatics Research, Stanford University, CA 94305; Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea, 443-480; Lister Hill National Center for Biomedical Communications (National Library of Medicine), National Institutes of Health, Bethesda, MD 20894; Department of Biostatistics, University of California, Los Angeles, CA 90095; Department of Human Genetics, University of California, Los Angeles, CA 90095; Real World Evidence Solutions, IMS Health, Burlington, MA 01803; Department of Medicine, University of Colorado School of Medicine, Aurora, CO 80045; Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37212; Geriatric Research, Education and Clinical Center, VA Tennessee Valley Healthcare System, Nashville, TN 37212; Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90089; Department of Pediatrics, University of Southern California, Los Angeles, CA 90089; Division of Health Sciences, University of South Australia, Adelaide, SA, Australia 5001; and Department of Statistics, Columbia University, New York, NY 10027

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
Population-level heterogeneity across systems, and patient-level heterogeneity within systems.
Conclusions: Network research

• It is feasible to encode the world population in a single data model

• Generating evidence is feasible
• Stakeholders willing to share results
• Able to accommodate vast differences in privacy and research regulation
• Incidence of side effects
• Any drug on the world market
• Any condition
• Absolute risk
  • Not causal (Characterization)
• On the Internet
OHDSI in Action

• Population-level estimation
What is the quality of the current evidence from observational analyses?

August 2010: “Among patients in the UK General Practice Research Database, the use of oral bisphosphonates was not significantly associated with incident esophageal or gastric cancer.”

Sept 2010: “In this large nested case-control study within a UK cohort [General Practice Research Database], we found a significantly increased risk of oesophageal cancer in people with previous prescriptions for oral bisphosphonates.”
Standard error vs effect size

Statistically significant
Observational research results in literature

85% of exposure-outcome pairs have p < 0.05

29,982 estimates
11,758 papers
Addressing reproducibility

1. Propensity stratification with *systematic* variable selection: measured confounding

2. Confidence interval calibration using negative controls: unmeasured confounding
3. **Multiple** databases, locations, practice types

4. Publish **all** hypotheses, code, parameters, runs
Addressing reproducibility

5. Carry out on aligned hypotheses at scale

<table>
<thead>
<tr>
<th>Amtriptiline</th>
<th>Bupropion</th>
<th>Citloproam</th>
<th>Desvenlafaxine</th>
<th>Doxepin</th>
<th>Duloxetine</th>
<th>Ocoonvulsive th</th>
<th>Escitalopram</th>
<th>Fluoxetine</th>
<th>Mirtazapine</th>
<th>Nortriptyline</th>
<th>Paroxetine</th>
<th>Psychotherapy</th>
<th>Sertraline</th>
<th>Venlafaxine</th>
<th>Vilazodone</th>
<th>Trazodone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Amtriptiline
- Bupropion
- Citloproam
- Desvenlafaxine
- Doxepin
- Duloxetine
- Ocoonvulsive th
- Escitalopram
- Fluoxetine
- Mirtazapine
- Nortriptyline
- Paroxetine
- Psychotherapy
- Sertraline
- Venlafaxine
- Vilazodone
- Trazodone
Estimates are in line with expectations

11% of exposure-outcome pairs have calibrated p < 0.05
<table>
<thead>
<tr>
<th>Class</th>
<th>Drug</th>
<th>Usual Dose, Range (mg/d)*</th>
<th>Daily Frequency</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary agents</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thiazide or thiazide-type diuretics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chlorthalidone</td>
<td>12.5–25</td>
<td>1</td>
<td>- Chlorthalidone is preferred on the basis of prolonged half-life and proven trial reduction of CVD.</td>
</tr>
<tr>
<td></td>
<td>Hydrochlorothiazide</td>
<td>25–50</td>
<td>1</td>
<td>- Monitor for hyponatremia and hypokalemia, uric acid and calcium levels.</td>
</tr>
<tr>
<td></td>
<td>Indapamide</td>
<td>1.25–2.5</td>
<td>1</td>
<td>- Use with caution in patients with history of acute gout unless patient is on uric acid–lowering therapy.</td>
</tr>
<tr>
<td></td>
<td>Metolazone</td>
<td>2.5–10</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ACE inhibitors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Benazepril</td>
<td>10–40</td>
<td>1 or 2</td>
<td>- Do not use in combination with ARBs or direct renin inhibitor.</td>
</tr>
<tr>
<td></td>
<td>Captopril</td>
<td>12.5–150</td>
<td>2 or 3</td>
<td>- There is an increased risk of hyperkalemia, especially in patients with CKD or in those on K⁺ supplements or K⁺-sparing drugs.</td>
</tr>
<tr>
<td></td>
<td>Enalapril</td>
<td>5–40</td>
<td>1 or 2</td>
<td>- There is a risk of acute renal failure in patients with severe bilateral renal artery stenosis.</td>
</tr>
<tr>
<td></td>
<td>Fosinopril</td>
<td>10–40</td>
<td>1</td>
<td>- Do not use if patient has history of angioedema with ACE inhibitors.</td>
</tr>
<tr>
<td></td>
<td>Lisinopril</td>
<td>10–40</td>
<td>1</td>
<td>- Avoid in pregnancy.</td>
</tr>
<tr>
<td></td>
<td>Moexipril</td>
<td>7.5–30</td>
<td>1 or 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Perindopril</td>
<td>4–16</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quinapril</td>
<td>10–80</td>
<td>1 or 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ramipril</td>
<td>2.5–10</td>
<td>1 or 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trandolapril</td>
<td>1–4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ARBs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Azilsartan</td>
<td>40–80</td>
<td>1</td>
<td>- Do not use in combination with ACE inhibitors or direct renin inhibitor.</td>
</tr>
<tr>
<td></td>
<td>Candesartan</td>
<td>8–32</td>
<td>1</td>
<td>- There is an increased risk of hyperkalemia in CKD or in those on K⁺ supplements or K⁺-sparing drugs.</td>
</tr>
<tr>
<td></td>
<td>Eprosartan</td>
<td>600–800</td>
<td>1 or 2</td>
<td>- There is a risk of acute renal failure in patients with severe bilateral renal artery stenosis.</td>
</tr>
<tr>
<td></td>
<td>Irbesartan</td>
<td>150–300</td>
<td>1</td>
<td>- Do not use if patient has history of angioedema with ARBs. Patients with a history of angioedema with an ACE inhibitor can receive an ARB beginning 6 weeks after ACE inhibitor is discontinued.</td>
</tr>
<tr>
<td></td>
<td>Losartan</td>
<td>50–100</td>
<td>1 or 2</td>
<td>- Avoid in pregnancy.</td>
</tr>
<tr>
<td></td>
<td>Olmesartan</td>
<td>20–40</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Telmisartan</td>
<td>20–80</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Valsartan</td>
<td>80–320</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>CCB—dihydropyridines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amlodipine</td>
<td>2.5–10</td>
<td>1</td>
<td>- Avoid use in patients with HFrEF; amlodipine or felodipine may be used if required.</td>
</tr>
<tr>
<td></td>
<td>Felodipine</td>
<td>5–10</td>
<td>1</td>
<td>- They are associated with dose-related pedal edema, which is more common in women than men.</td>
</tr>
<tr>
<td></td>
<td>Isradipine</td>
<td>5–10</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nicardipine SR</td>
<td>5–20</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nifedipine LA</td>
<td>60–120</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nisoldipine</td>
<td>30–90</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>CCB—nondihydropyridines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diltiazem SR</td>
<td>180–360</td>
<td>2</td>
<td>- Avoid routine use with beta blockers because of increased risk of bradycardia and heart block.</td>
</tr>
<tr>
<td></td>
<td>Diltiazem ER</td>
<td>120–480</td>
<td>1</td>
<td>- Do not use in patients with HFrEF.</td>
</tr>
<tr>
<td></td>
<td>Verapamil IR</td>
<td>40–80</td>
<td>3</td>
<td>- There are drug interactions with diltiazem and verapamil (CYP3A4 major substrate and moderate inhibitor).</td>
</tr>
<tr>
<td></td>
<td>Verapamil SR</td>
<td>120–480</td>
<td>1 or 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Verapamil-delayed onset ER (various forms)</td>
<td>100–480</td>
<td>1 (in the evening)</td>
<td></td>
</tr>
</tbody>
</table>
Evidence to support the guideline

- 40 randomized trials
- Most decisions are “expert opinion”
## Comparisons of hypertension treatments

<table>
<thead>
<tr>
<th>Type of Comparison</th>
<th>Theoretical</th>
<th>Observed (n &gt; 2,500)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single ingredients</td>
<td>58</td>
<td>39</td>
</tr>
<tr>
<td>Single ingredient comparisons</td>
<td>58 * 57 = 3,306</td>
<td>1,296</td>
</tr>
<tr>
<td>Single drug classes</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>Single class comparisons</td>
<td>15 * 14 = 210</td>
<td>156</td>
</tr>
<tr>
<td>Dual ingredients</td>
<td>58 * 57 / 2 = 1,653</td>
<td>58</td>
</tr>
<tr>
<td>Single vs duo drug comparisons</td>
<td>58 * 1,653 = 95,874</td>
<td>3,810</td>
</tr>
<tr>
<td>Dual classes</td>
<td>15 * 14 / 2 = 105</td>
<td>32</td>
</tr>
<tr>
<td>Single vs duo class comparisons</td>
<td>15 * 105 = 1,575</td>
<td>832</td>
</tr>
<tr>
<td>Duo vs duo drug comparisons</td>
<td>1,653 * 1,652 = 2,730,756</td>
<td>2,784</td>
</tr>
<tr>
<td>Duo vs duo class comparisons</td>
<td>105 * 104 = 10,920</td>
<td>992</td>
</tr>
<tr>
<td>Total comparisons</td>
<td>2,843,250</td>
<td>10,278</td>
</tr>
<tr>
<td>Outcomes of interest</td>
<td>58</td>
<td>58</td>
</tr>
<tr>
<td>Target-comparator-outcomes</td>
<td>2,843,250 * 58 = 164,908,500</td>
<td>587,020</td>
</tr>
</tbody>
</table>
Cardiovascular efficacy by drug

Prescriptions are not written at the class-level; must choose an individual drug for the patient

- $1^{st}$-line > $2^{nd}$-line
- Some within-class differences failed diagnostics, e.g. captopril

Composite (MI, HF, stroke) outcome in meta-analysis
HEAD-to-HEAD HTN drug comparisons

- Trials: 40
- $N = 102 - [1148] - 33K$
- Comparisons: 10,278
- $N = 3502 - [212K] - 1.9M$
OHDSI in Action

• Patient-level prediction
Patient-level prediction
Stroke risk in atrial fibrillation

The OHDSI approach lets the model choose from all conditions and drugs

247 variables out of 16900 including:
1. all the CHADS2 (afib stroke risk) markers
2. plus some other variables that make clinical sense (ex: brain cancer, smoking)
3. plus some other variables that warrant further exploration (ex: antiepileptic, COPD)
<table>
<thead>
<tr>
<th>AMI</th>
<th>Acute heart injury</th>
<th>Alpea</th>
<th>Constipation</th>
<th>Decreased libido</th>
<th>Diarrhea</th>
<th>Hypothyroidism</th>
<th>Nausea</th>
<th>Stroke</th>
<th>Vomiting</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>1.00</td>
<td>0.90</td>
<td>0.80</td>
<td>0.70</td>
<td>0.60</td>
<td>0.50</td>
<td></td>
<td>OPTUM</td>
<td>MDCD</td>
<td>MDCR</td>
</tr>
<tr>
<td></td>
<td>Gradient boosting</td>
<td>Random forest</td>
<td>Regularized regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Transportability Assessment Stroke

Transportability to MDCR is low
Conclusions

• It is feasible to create an enormous international research network
• Sites will volunteer to run studies
• Completely open
  – Data model, methods, tools
• Concrete approach to address the credibility crisis
• Prediction
  – It’s not ROC area
  – New, useful information
Join the journey

http://ohdsi.org