An Open Science Approach to Medical Evidence Generation: Introducing Observational Health Data Sciences and Informatics

Jon Duke, MD MS
Regenstrief Institute
Academy Heath
June 14 2015
What is OHDSI?

• The Observational Health Data Sciences and Informatics (OHDSI) program is a multi-stakeholder, interdisciplinary collaborative

• The goal of OHDSI is to bring out the value of observational health data through large-scale analytics and evidence generation

• All our software and other products are released as open-source

http://ohdsi.org
OHDSI: a global community

OHDSI Collaborators:
- >140 researchers in academia, industry and government
- >10 countries

OHDSI Data Network:
- >50 databases standardized to OMOP common data model
- >680 million patients
OHDSI Evidence Generation

• Clinical characterization:
  – Descriptive statistics (e.g., natural history of a disease or patterns of medication use)
  – Quality improvement (e.g., performance measures)

• Population-level estimation
  – Safety surveillance (e.g., identifying new adverse event risks for drugs)
  – Comparative effectiveness (e.g., comparing interventional to non-interventional treatment of chronic back pain)

• Patient-level prediction
  – Incorporating patient medical history to provide personalized recommendations for therapy selection, adverse event risk, high value diagnostic studies
The odyssey to evidence generation
Open Science through Standardization

• The OHDSI community has standardized core components of the research process in order to
  – Promote transparent, reproducible science
  – Reveal data quality issues
  – ‘Calibrate’ datasets
  – Bring skillsets together from across the community (clinical, epi, stats, compSci)
Opportunities for standardization in the evidence generation process

- **Data structure**: tables, fields, data types
- **Data content**: vocabulary to codify clinical domains
- **Data semantics**: conventions about meaning
- **Cohort definition**: algorithms for identifying the set of patients who meet a collection of criteria
- **Covariate construction**: logic to define variables available for use in statistical analysis
- **Analysis**: collection of decisions and procedures required to produce aggregate summary statistics from patient-level data
- **Results reporting**: series of aggregate summary statistics presented in tabular and graphical form
One model, multiple use cases
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 tools built to help

OHDSI Forums:
Public discussions for OMOP CDM Implementers/developers

http://github.com/OHDSI
Standardized large-scale analytics tools under development within OHDSI

ACHILLES: Database profiling
CIRCE: Cohort definition
HERMES: Vocabulary exploration

CALYPSO: Feasibility assessment

HERACLES: Cohort characterization

OHDSI Methods Library: CYCLOPS
CohortMethod
SelfControlledCaseSeries
SelfControlledCohort
TemporalPatternDiscovery
Empirical Calibration

LAERTES: Drug-AE evidence base

PLATO: Patient-level predictive modeling
HOMER: Population-level causality assessment

Patient-level data in OMOP CDM

http://github.com/OHDSI
OHDSI Software

- Community developed
- Apache 2.0 licensed
- Available on GitHub
- Common frameworks
  - Java
  - HTML5 / Javascript
  - R
  - Oracle / SQL Server / Postgres / Redshift / Netezza
Motivating example to see the OHDSI tools in action

MINI-SENTINEL MEDICAL PRODUCT ASSESSMENT
A PROTOCOL FOR ASSESSMENT OF DABIGATRAN

Version 3

March 27, 2015
Prior versions:
Version 1: December 31, 2013
Version 2: March 18, 2014

Prepared by: Alan S. Go, MD¹, Daniel Singer, MD², T. Craig Cheetham, PharmD MS³, Darren Toh, ScD⁴, Marsha Reichman, PhD⁵, David Graham, MD MPH⁶, Mary Ross Southworth, PharmD⁶, Rongmei Zhang PhD⁷, Monika Houstoun, PharmD⁵, Yu-te Wu PhD⁷, Katrina Mott MS⁵, Joshua Gagne, PharmD ScD⁸

Author Affiliations: 1. Division of Research, Kaiser Permanente Northern California, Oakland, CA. 2. General Medicine Division, Massachusetts General Hospital, Boston, MA. 3. Kaiser Permanente Southern California, Downey, CA. 4. Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA. 5. Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research (CDER), Food and Drug Administration (FDA), Silver Spring, MD. 6. Division of Cardiovascular and Renal Products, Office of New Drugs, Center for Drug Evaluation and Research (CDER), Food and Drug Administration (FDA), Silver Spring, MD. 7. Division of Biometric VII, Office of Biostatistics, Office of Translation Sciences, Food and Drug Administration (FDA), Silver Spring, MD. 8. Division of ...
III. PROTOCOL DETAILS

A. ASSESSMENT DESIGN
This one-time assessment will employ a “new user” parallel cohort design.¹²

B. COHORT IDENTIFICATION

1. Target Population
We will focus on the identification of adult (age ≥21 years) patients with diagnosed nonvalvular atrial fibrillation and who are new users of dabigatran or warfarin.

2. Sample Inclusion and Exclusion Criteria
The target sample inclusion and exclusion criteria are summarized in Table 1 below. Please see Appendix A and Section D for additional details, definitions and rationale.

Table 1. Inclusion and exclusion criteria for comparison of adults with atrial fibrillation who are new users of dabigatran or warfarin in the MSDD.

<table>
<thead>
<tr>
<th>Inclusion Criteria</th>
<th>Exclusion Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>• First dispensing of dabigatran or warfarin therapy from November 1, 2010 to the</td>
<td>• Less than 180 days of continuous enrollment with prescription and medical</td>
</tr>
<tr>
<td>the most recent data available in the MSDD from participating Data Partners</td>
<td>coverage immediately preceding the date of the index dispensing (i.e., index date)</td>
</tr>
<tr>
<td>• Age 21 years or older at the first dispensing of dabigatran or warfarin therapy</td>
<td>• Any prior dispensing for warfarin, dabigatran, rivaroxaban or apixaban during the</td>
</tr>
<tr>
<td>• One or more diagnoses of atrial fibrillation or atrial flutter based on ICD-9-CM</td>
<td>180 days before index date**</td>
</tr>
<tr>
<td>codes (ICD-9-CM 427.31, 427.32) from any practice setting (inpatient or outpatient)</td>
<td>• Known mechanical heart valve or diagnosed mitral stenosis at index date based on</td>
</tr>
<tr>
<td>any time before the first identified prescription for dabigatran or warfarin</td>
<td>corresponding administrative diagnosis and/or procedure codes</td>
</tr>
<tr>
<td>therapy during the study period*</td>
<td>• Chronic hemodialysis or peritoneal dialysis at index date based on corresponding</td>
</tr>
<tr>
<td></td>
<td>administrative diagnosis and/or procedure codes</td>
</tr>
<tr>
<td></td>
<td>• History of kidney transplant at index date based on corresponding administrative</td>
</tr>
<tr>
<td></td>
<td>diagnosis and/or procedure codes</td>
</tr>
<tr>
<td></td>
<td>• At a skilled nursing facility or nursing home at index date</td>
</tr>
</tbody>
</table>
Let’s ask the OHDSI network!
OLYMPUS
THE OHDSI APPLICATION LAUNCHER

There are remote WebAPIs configured. Applications that support toggling between WebAPIs will allow you to use these via the gear/settings.

ATHENA
OMOP Vocabulary Loader

CIRCE
Cohort Creation

HERMES
OMOP Vocabulary Explorer

ACHILLES
Dataset Characterization

HERACLES
Cohort Characterization

CALYPSO
Clinical Trial Feasibility
Use ACHILLES to see if the databases have the required data elements
Also use ACHILLES to check for any data quality issues

<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>
Use HERMES to figure out how to find a particular condition, drug, procedure, or other concept.
Use CIRCE to define the cohort of interest

Index Population: MiniSentinel replication - warfarin new users

Description:

Expression

People having any of the following: Add Primary Event Filters...

- a drug era of warfarin
- for the first time in the person’s history
- era start is: After 2010-11-01
- with age at era start Greater or Equal To 21

with observation at least 180 days prior and 0 days after index

Limit primary events to: All Events per person.
Use CALYPSO to conduct feasibility assessment to evaluate the impact of study inclusion criteria.
Use HERACLES to characterize the cohorts you developed.
Use HERACLES to characterize the cohorts you developed.
Use LAERTES to summarize evidence from existing data sources
Step up to Advanced Analytic Methods

https://github.com/OHDSI

Observational Health Data Sciences and Informatics

CohortMethod
An R package for performing new-user cohort studies in an observational database in the OMOP Common Data Model.
Updated 10 days ago

SelfControlledCohort
[Under development] Method to estimate risk by comparing time exposed with time unexposed among the exposed cohort
Updated on Dec 22, 2014

http://github.com/OHDSI
Why is this a novel approach?

- Large-scale analytics, scalable to ‘big data’ problems in healthcare:
  - millions of patients
  - millions of covariates
  - millions of questions

- End-to-end analysis, from CDM through evidence
  - No longer de-coupling ‘informatics’ from ‘statistics’ from ‘epidemiology’
Standardize Analysis and Results Reporting

```r
summary(outcomeModel)
```
Demo!

OLYMPUS
THE OHDSI APPLICATION LAUNCHER

There are remote WebAPIs configured. Applications that support toggling between WebAPIs will allow you to use these via the gear//settings.

ATHENA
OMOP Vocabulary Loader

CIRCE
Cohort Creation

HERMES
OMOP Vocabulary Explorer

HERACLES
Cohort Characterization

ACHILLES
Dataset Characterization

CALYPSO
Clinical Trial Feasibility
Concluding Thoughts

• Open science requires optimized technical infrastructure, community infrastructure, and dedication

• But open science is not charity!
  – The payoff can be both for individual participants and the community

• A diversity of skillsets brings value to all and greatly accelerates generation of high quality evidence
Join the journey

Interested in OHDSI?
Questions or comments?

Contact:

jonduke@regenstrief.org