Open-Source Big Data Analytics in Healthcare

Patrick Ryan\textsuperscript{a}, Jon Duke\textsuperscript{b}, George Hripcsak\textsuperscript{c}

\textsuperscript{a} Global Epidemiology, Janssen Research and Development, Titusville, NJ, USA
\textsuperscript{b} Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN, USA
\textsuperscript{c} Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract

In this tutorial, we will introduce attendees to an open-source platform for conducting large-scale analytics on observational health data, such as electronic medical records, administrative claims data, or other sources of patient information. Our focus will be on the use of the OMOP Common Data Model (CDM) and the Observational Health Data Sciences and Informatics (OHDSI) analytics platform. We will guide participants through 1) how to transform data sources into the standardized OMOP data model; 2) how to characterize population-level data for uses such as clinical benchmarking, research, and data quality; and 3) conducting analyses across a distributed network of data sites.

Keywords:
Data analytics, Observational Data, Common Data Model, OMOP, Observational Health Data Sciences and Informatics.

Tutorial Description

This tutorial will enable participants to leverage a rapidly growing set of open-source analytics tools designed for use with health care data. Longitudinal observational patient-level data offer tremendous potential to explore clinically relevant questions requiring reliable scientific evidence about disease natural history, healthcare delivery, and the effects of medical interventions through population-level estimation such as medical product surveillance and comparative effectiveness research and patient-level predictions. Applying a data science paradigm to realize this potential requires data holders to establish cross-disciplinary expertise in medical informatics, statistics, and epidemiology in order to follow the full process from source patient-level data through transformation to a common data model with standardized vocabulary through aggregate summary statistics through interactive exploratory visualization of results and clinical interpretation of research findings. The OHDSI framework provides analysis tools and standard processes to achieve this aim.

At the conclusion of this tutorial, participants will be able to transform their data into an internationally used standard model, conduct a wide range of analyses and explorations of their own data, and participate in collaborative analyses across a large network of sites.

Part One

Steps to Standardized Data

Data standardization is the process of bringing diverse data into a common format that allows for collaborative research, large-scale analytics, and sharing of sophisticated tools and methodologies. The OMOP Common Data Model was first developed in 2008 and has grown into a widely used standard model for representing healthcare data. The process of converting a dataset into a new model is often referred to as an ETL process (Extract-Transform-Load). In this first part of the tutorial, we will describe the ETL process for moving data to the OMOP CDM. Fortunately, this process is facilitated by several community-developed tools that we will demonstrate. Shown in Figure 1 is a screenshot of the WhiteRabbit tool, a software package that analyzes a source dataset, determines characteristics of its tables and content, and facilitates mapping to the equivalent tables and fields in the OMOP CDM.

In addition to transforming to a common data structure, the OMOP CDM provides facilities to standardize data content to a common set of vocabularies. Community standard vocabularies have been selected for all clinical domains (conditions, drugs, procedures, measurements) and source codes can be mapped into these standards are part of the ETL
process. OHDSI collaborators have developed a tool, Usagi, shown in Figure 2, which enables users to map source codes that appear in the patient-level data to standard concepts. Together with the standard vocabulary, this tool has enabled organizations around the world with local or non-standard codes to be combined into a common platform.

Figure 2. Screenshots of the open-source USAGI vocabulary mapping tool.

Part Two
Exploring Population-Level Data

Once data are transformed into the OMOP CDM, a wide range of advanced tools are available for data exploration, cohort generation, research, care quality, and other data-driven initiatives. The OHDSI collaborative has produced a suite of open-source tools designed to support these use cases. In this tutorial we will discuss mature user-friendly tools such as ACHILLES (Figure 3). ACHILLES is a platform which enables the characterization, quality assessment and visualization of observational health databases. ACHILLES provides users with an interactive, exploratory framework to assess patient demographics, the prevalence of conditions, drugs and procedures, and to evaluate the distribution of values for clinical observations. ACHILLES is intended to be implemented by organizations that have patient-level observational health databases available in their local environment. By itself, ACHILLES does not perform study-specific analysis, but can assist exploring the contents of a CDM database to determine whether data exists that can support a study. We will also discuss opportunities for more advanced users with familiarity with SQL and analytics packages such as R.

Part Three
Performing a Network Analysis

Often, an individual organizations do not have a sufficient volume of patients with a given condition to generate the power necessary for a research study. The OHDSI Research Network provides a means to conduct studies at multiple sites using a common platform. In this section, we will show how studies can be conducted across this network.

Study Phase One: Design

OHDSI uses a standardized proposal mechanism (Figure 4) in which a the lead researcher describes the study objectives and protocol. The proposer solicits feedback from other network collaborators to ensure the design meets community standards. The lead researcher also uses this review period to generate interest from other researchers.

Figure 4. Example of an OHDSI Network research protocol.

Treatment Pathways in Chronic Disease

Objective: The objective of this study is to characterize the prevalence of different treatment pathways for three chronic diseases: Hypertension, Type II Diabetes, and Depression. We will systematically summarize the treatment pathways observed among patients who have at least 3 years of continuous observation and persistent treatment following initiation. We will stratify the results by year to evaluate temporal trends, and will further stratify by data source to determine if treatment pathways vary by population, geography, and data capture process.

Rationale: While numerous treatment guidelines exist for chronic conditions, there is a paucity of data on the real-world treatment pathways that patients experience in practice. Understanding these pathways is essential for establishing context around questions of drug utilization, effectiveness, and adherence.

Project Leads: Patrick Ryan, Jon Duke, George Hrozencik, Martha Schumere, Nigam Sheth

Coordinating Institution(s): Janssen R&D, Columbia University, Regenstrief Institute, Stanford University

Additional Participants:

Full Protocol: Hypertension Treatment Pathways 12-4-2014
Initial Proposal Date: 12:35/2014
Launch Date: 12:35/2014
Study Closure Date: 12:31/2014
Results Submission: Email or STTP
Study Phase Two: Standardized Analysis Code

A highly valuable aspect of research conducted via the OMOP CDM is that a common code base can be used across multiple institutions with varying data models. The OHDSI Research Network provides an additional technology layer to translate across multiple database dialects (e.g., SQL Server, Oracle, Postgres). This generalizability greatly increases the range of organizations that can participate in studies. In order to support this functionality, the lead researcher must generate analysis code (typically using SQL and/or R) that will be shared with all participating data sites.

Study Phase 3: Results and Dissemination

Once the study code is complete, all participating sites run the code locally on their own data. The resulting output (comprising only aggregate, de-identified information) is forwarded to the study coordinator, who is responsible of synthesizing and disseminating these results. Dissemination may incude data visualization (example in Figure 5). All OHDSI study methods and results are made publically available.

Figure 5. Results of OHDSI study on treatment pathways in hypertension.

Summary

In this tutorial, participants will gain hands-on experiences using a robust open-source platform for data analytics in healthcare. They will gain direct experience in leveraging a large Research Network to conduct multi-site observational studies.

References


