MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)
© 2015 IMIA and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-564-7-574

Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers

George Hripcsak[a], Jon D. Duke[b], Nigam H. Shah[c], Christian G. Reich[d], Vojtech Huser[e], Martijn J. Schuemie[f,g], Marc A. Suchard[h], Rae Woong Park[i], Ian Chi Kei Wong[f], Peter R. Rijnbeek[j], Johan van der Lei[j], Nicole Pratt[k], G. Niklas Norén[l], Yu-Chuan Li[m], Paul E. Stang[g], David Madigan[n], Patrick B. Ryan[g]

[a] *Department of Biomedical Informatics, Columbia University Medical Center, New York, NY, USA*
[b] *Regenstrief Institute, Indianapolis, IN, USA*
[c] *Center for Biomedical Informatics Research, Stanford University, CA, USA*
[d] *AstraZeneca PLC, Waltham, MA, USA*
[e] *NIH Clinical Center, Bethesda, MD, USA*
[f] *Centre for Safe Medication Practice and Research, Dept. of Pharmacology and Pharmacy, University of Hong Kong, Hong Kong*
[g] *Janssen Research & Development, LLC, Titusville, NJ, USA*
[h] *Dept. of Biomathematics & Dept. of Human Genetics, David Geffen School of Medicine, Uni. of California, Los Angeles, CA, USA*
[i] *Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea*
[j] *Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands*
[k] *School of Pharmacy and Medical Sciences, University of South Australia, Australia*
[l] *Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring, Uppsala, Sweden*
[m] *College of Medical Science and Technology (CoMST), Taipei Medical University, Taipei, Taiwan*
[n] *Department of Statistics, Columbia University, New York, NY, USA*

Abstract

The vision of creating accessible, reliable clinical evidence by accessing the clincial experience of hundreds of millions of patients across the globe is a reality. Observational Health Data Sciences and Informatics (OHDSI) has built on learnings from the Observational Medical Outcomes Partnership to turn methods research and insights into a suite of applications and exploration tools that move the field closer to the ultimate goal of generating evidence about all aspects of healthcare to serve the needs of patients, clinicians and all other decision-makers around the world.

Keywords:

Health Services Research; Databases; Observation.

Introduction

Observational Health Data Sciences and Informatics (OHDSI, pronounced "Odyssey") [1] is an international collaborative whose goal is to create and apply open-source data analytic solutions to a large network of health databases to improve human health and wellbeing. The OHDSI team comprises academics, industry scientists, health care providers, and regulators whose formal mission is to transform medical decision making by creating reliable scientific evidence about disease natural history, healthcare delivery, and the effects of medical interventions through large-scale analysis of observational health databases for population-level estimation and patient-level predictions [2]. Over 90 participants from around the world have joined the collaborative with a vision to access a network of one billion patients to generate evidence about all aspects of healthcare, where patients, clinicians and all other decision-makers around the world use OHDSI tools and evidence every day [3].

Methods

OHDSI grew out of the Observational Medical Outcomes Partnership (OMOP) [4], which was a public-private partnership established in the US to inform the appropriate use of observational healthcare databases for studying the effects of medical products. The five-year project developed new methods in observational research and established an observational research laboratory. At the conclusion of this five-year project, the OMOP research investigators initiated the OHDSI effort. The research laboratory moved to the Reagan-Udall Foundation for the FDA under the Innovation in Medical Evidence Development and Surveillance (IMEDS)

Table 1. Tables in the OMOP Common Data Model V5.0

Model Domain	Table Names
Standardized Clinical Data Tables	PERSON, OBSERVATION_PERIOD, SPECIMEN, DEATH, VISIT_OCCURRENCE, PROCEDURE_OCCURRENCE, DRUG_EXPOSURE, DEVICE_EXPOSURE, CONDITION_OCCURRENCE, MEASUREMENT, NOTE, OBSERVATION, FACT_RELATIONSHIP
Standardized Health System Data Tables	LOCATION, CARE_SITE, PROVIDER
Standardized Health Economics Data Tables	PAYER_PLAN_PERIOD, VISIT_COST, PROCEDURE_COST, DRUG_COST, DEVICE_COST
Standardized Derived Elements	COHORT, COHORT_ATTRIBUTE, DRUG_ERA, DOSE_ERA, CONDITION_ERA

Program [5].

A centerpiece of the OMOP project was the development of the OMOP Common Data Model (CDM) [4] which represents healthcare data from diverse sources in a consistent and standardized way (see Table 1). This CDM is a "strong" information model, in which the encoding and relationships among concepts are explicitly and formally specified. The OHDSI team has adopted and continued maintenance of this model and its associated vocabulary services. OHDSI's overall approach is to create an open network of observational data holders, and require that they translate their data to the OMOP CDM. Each element in the participant's database must be mapped to the approved CDM vocabulary and placed in the data schema. In return, this approach creates a unique opportunity of implementing a number of existing data exploration and evidence generation tools and participating in world-wide studies because any given query can be executed at any site without modification. This enables multicenter, global analyses to be executed rapidly and efficiently using applications or programs developed at a single site.

Data are retained at the participant's site, simplifying patient and business privacy issues. The team previously found that simply merging the databases is likely to give poor answers because of heterogeneity [6]. Instead, analyses are carried out locally and the results transmitted to the coordinating center, where they can be studied on a population level and aggregated as appropriate.

OHDSI operates at several levels: infrastructure, data, methods, applications, and experiments. These levels serve both to support and inform the work of each other to ensure that the infrastructure and products support the mission. Rather than just creating a data network, OHDSI directly integrates researchers who use the network and data scientists who create the algorithms with the use cases for the data network.

The group's guiding principles are that the effort be:

1. Evidence-based, such that OHDSI's scientific research and development are driven by objective, empirical evidence to ensure accuracy and reliability;
2. Practical, going beyond methodological research, but developing applied solutions and generating clinical evidence;
3. Comprehensive, aiming to generate reliable scientific evidence for all interventions and all outcomes;
4. Transparent, such that all work products within OHDSI are Open Source and publicly available, including source code, analysis results, and other evidence generated in all our activities;
5. Inclusive, encouraging active participation from all stakeholders – patients, providers, payers, government, industry, academia – in all phases of research and development; and finally
6. Secure, protecting patient privacy and respecting data

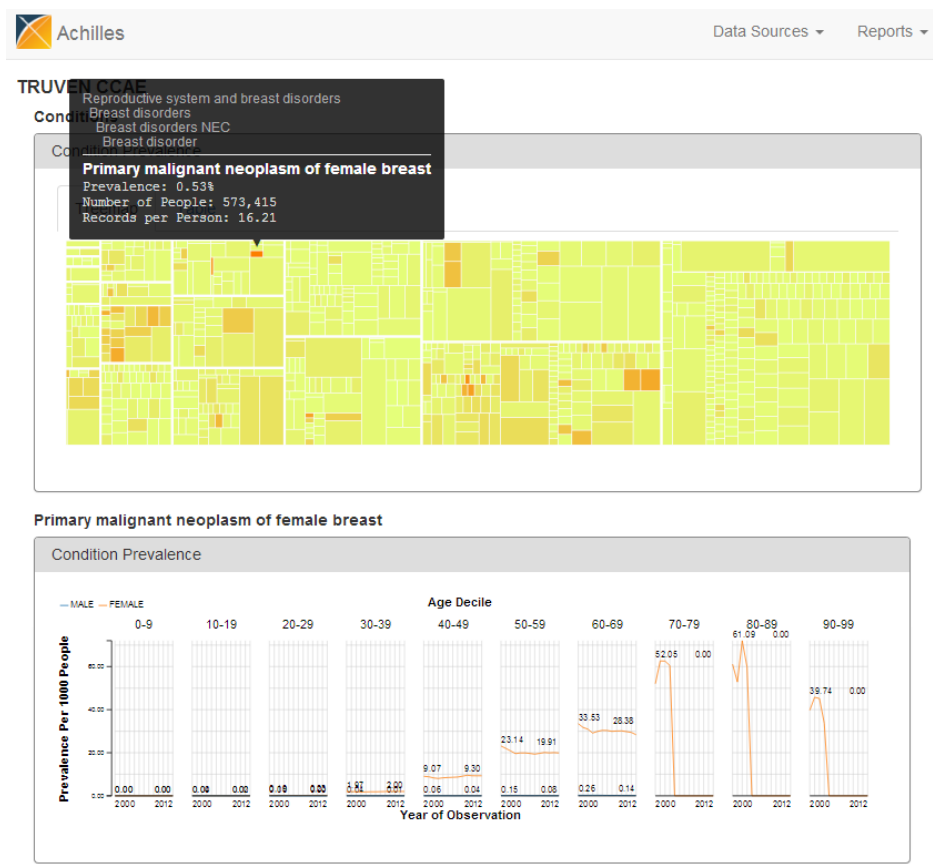

Figure 1. ACHILLES. A tree map (top half) summarizes the content of the database, where the rectangle size indicates prevalence and color indicates average number of records per patient. The bottom half reports prevalence by year, age, and sex.

holder interests at all times.

To achieve the principle of inclusivity, OHDSI is an open collaborative. Anyone who can give time, data, or funding is welcome, and participation in the operation of OHDSI is expected. Currently, participants come from around the world, including the United States, United Kingdom, Netherlands, Sweden, Italy, Korea, Taiwan, Hong Kong, and Australia.

Software Tools

OHDSI is building a suite of tools to faciliate data exploration and evidence generation. The first of these, ACHILLES (Automated Characterization of Health Information at Large-scale Longitudinal Exploration System), is a browser-based visualization tool for reviewing a clinical database based on pre-extracted summary statistics from datasets in OMOP CDM format. ACHILLES enables characterization, quality assessment, and visualization of observational data; and provides users with an interactive, exploratory framework to assess patient demographics and the prevalence of all conditions, drugs, procedures, and observations stored in the dataset. The ACHILLES application and source code is available in the public domain [6] and a demonstration is hosted on the OHDSI Web site [7].

ACHILLES has two main components. The first component is implemented as an R package and runs securely within an organization's local environment without disclosing any patient identifiable information. The R package generates summary statistics that describe the quality and content of the patient-level observational health database and provides features to perform a simple review or bulk export of the summary statistics in JSON data files. The second component of ACHILLES is implemented as an HTML5 / JavaScript website with a series of interactive reports that allow exploration and visualization of generated summary statistics. Summary statistics from multiple databases can be made available from a single installation of the ACHILLES website.

Data owners have used ACHILLES to assess the quality of their database, looking for gaps that may signify upload errors. Other investigators have used ACHILLES to do an initial assessment of whether the database is likely to hold a sufficient number of cases of interest to be worth investigating further. Figure 1 shows a typical visualization for a researcher interested in primary malignant neoplasm of the female breast. As shown, ACHILLES displays the prevalence of the condition, the depth of data on those patients, the age distribution, the sex distribution (less relevant here), and the time of observation. Other current ACHILLES views show the temporal characteristics of the data (e.g., prevalence by month), quality reports, and treemaps of subsets of data where each box's size and color represent different database metrics. Its data quality reports help the data owner curate the database and allow outsiders to review supported aspects of data quality.

Additional OHDSI tools are in development. HERMES (Health Entity Relationship and Metadata Exploration System) is a web-based vocabulary browsing tool with the ability to search for a term and explore related concepts. PLATO (Patient-Level Assessment of Treatment Outcomes) provides predictive models that assess probability of a patient experiencing any outcome following initiation of any intervention, given his or her personal medical history. For example, a patient could enter his gender, age, primary diagnosis, and medication to check the prevalence of side effects to the medication. HERACLES (Health Enterprise Resource and Care Learning Exploration System) helps the user to build and explore cohorts to assess a specific clinical population across a wide-variety of clinical dimensions, including specialized analytics for performing clinical quality metrics. And HOMER (Health Outcomes and Medical Effectiveness Research) enables risk identification and comparative effectiveness studies, with real-time exploration of the effects of medical products. HOMER supports exploratory analyses for a wide variety of dimensions that serve as evidence for or against causality, such as the

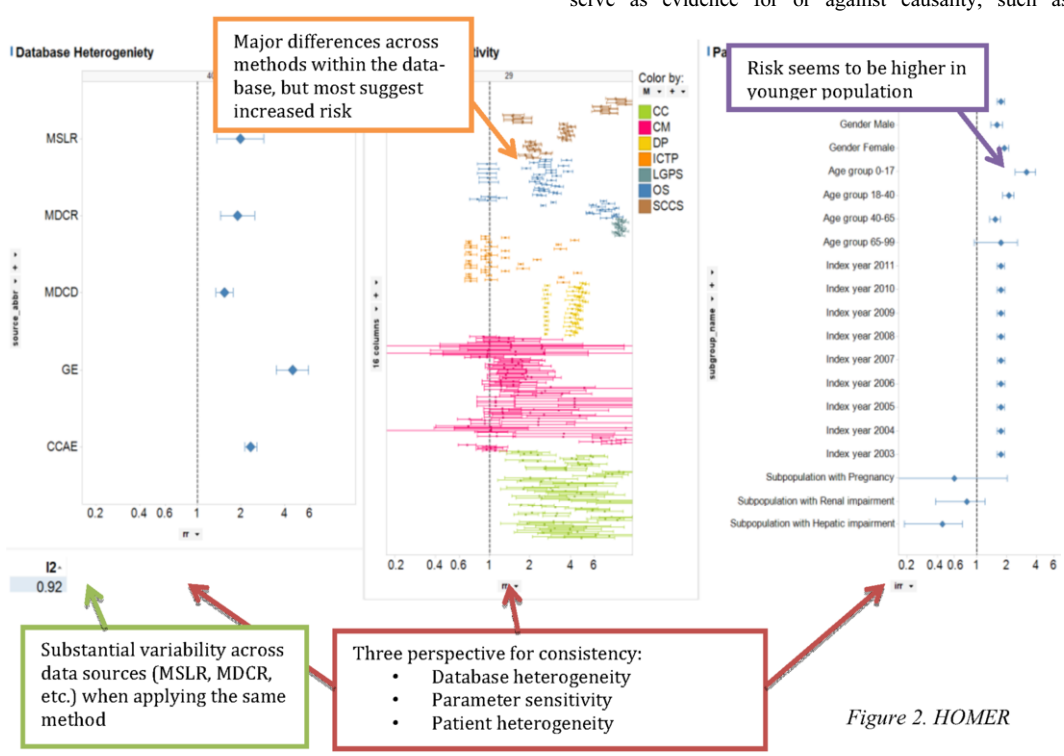

Figure 2. HOMER

consistency of findings. Consistency can be assessed across source datasets, analysis methods, or patient subpopulations. Figure 2 shows a prototype screen from HOMER, in which the relationship between a drug and adverse effect are explored across multiple dimensions.

In addition to creating user-friendly software, OHDSI also engages in the development of advanced analytic methods, including making Bayesian regression feasible over large data sets, handling sparse coding, and performing hierarchical association mining. Finally, OHDSI has been a leader in application of empircal calibration techniques to healthcare datasets to account for the biases inherent in these data sources as another contribution to the transparency and integrity of observational data research.

Research Network

The goal of the Research Network is to lower barrier to performing large-scale collaborative research using observational data and to generate high-quality evidence through peer review across study design, execution, and data analysis. Data owners make data available in return for algorithms and tools and the ability to post queries to the network. OHDSI has developed a process for the development and execution of research studies across the data network. This process includes proposal by a collaborator, review by the community, and promotion to active project. Initial proposals typically comprise a basic protocol including objective, rationale, target population, and initial source code. Once posted to the OHDSI Research Forum [8], the proposed project will be reviewed by other collaborators to determine interest level as well as to generate suggestions for modification to the design, target phenotypes, or analytic methods. Projects that generate interest from multiple sites and produce a complete protocol and cross-tested code are promoted to an Active Project. All community members are invited to run the analysis on their local dataset and return the results (de-identified, aggregate) centrally. Analyses across several sites rely on CDM standardization of various data domains to common terminologies, such as SNOMED for diagnoses, RxNorm for drug ingredients and LOINC for laboratory results. OHDSI's choice of standard terminologies does not limit who may participate because mappings and tools are supplied to translate from other terminologies. After a defined period, result submission is closed and the data are analyzed and presented back to the community. Based on these findings, publications and follow-up studies may result.

OMOP previously demonstrated that databases from different sources can give vastly different answers [9]. A blind aggregation of the results may increase variance due to heterogeneity instead of decreasing it due to sample size. Therefore, depending on the research question and the participating sites, the collected results may simply be reported without aggregation, summarized, or, if sufficient homogeneity can be demonstrated, aggregated.

Results

OHDSI held its first annual meeting at Columbia University Medical Center in New York City on 16 and 17 of October 2014. Fifty-eight participants reviewed the vision and goals giving rise to the creation of OHDSI and formed working groups to address the common data model, vocabulary, knowledge bases, estimation methods, phenotype generation, clinical characterization, and cohort definition. The outcomes of the meeting included the following:

1. Confirmation of the commitment of data custodians to participate in federated research studies;
2. Decision to open the database to queries from external researchers under a formal process; and
3. Work progress on each of the working group areas.

In preparation for the meeting, the OHDSI team surveyed current users of the OMOP data model. They found that 58 existing OMOP databases have collectively converted 682 million patient records to the OMOP CDM. This large number includes both patients and sources that are duplicated across databases and also includes databases that are not currently participating in the OHDSI Research Forum. Nevertheless, the total count demonstrates the feasibility of imposing a strong information model and executing a CDM conversion on a number of records that would represent a significant fraction of the world's population. In other words, the OHDSI vision is, in fact, feasible today. Furthermore, we estimate that the actual number of patients currently available in the OHDSI Research Forum is over 200 million.

OHDSI has just begun to distribute research queries. The first published OHDSI study used databases maintained by one site to carry out a medication-wide association study, assessing whether drugs with similar function or structure cause similar side effects [10].

Discussion and Conclusion

We envision a future where observational studies will inform clinical practice—by providing practice-based evidence—using the unprecedented amount of available patient data and the use of computerized systems to process the data [11]. Feinstein et al. initiated the idea of using data on 678 lung cancer patients as an electronic 'library of clinical experience' to obtain a personalized prognosis [12]. We believe that it now should be possible to provide not only prognosis information but also provide estimates on the comparative effectiveness as well as patient level assesment of treatment options.

We are not alone in this endeavor. For example, PCORnet [13] is designed to improve the national infrastructure for conducting clinical outcomes research. The network will enable a national capacity to conduct comparative effectiveness research efficiently and to learn from the health care experiences of millions of Americans.

OHDSI is implementing such a collaboration internationally by building on the OMOP experience for observational research. The existence of multiple very large databases that use the OMOP Common Data Model demonstrates that a worldwide conversion of clinical data from all patients is in fact feasible. Previous success in algorithm development, software distribution, and evidence generation points to potential success in the overall evidence-generating OHDSI project.

Acknowledgments

This work was funded in part by US National Science Foundation grant NSF IIS 1251151.

References

[1] OHDSI. http://ohdsi.org/ Accessed 2014-12-07.
[2] OHDSI Vision. http://ohdsi.org/who-we-are/mission-vision-values/ Accessed 2014-12-07.
[3] OHDSI Values. http://ohdsi.org/who-we-are/mission-vision-values/ Accessed 2014-12-07.
[4] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety

surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60. Epub 2011 Oct 28.
[5] IMEDS. http://imeds.reaganudall.org/ Accessed 2014-12-07.
[6] ACHILLES Repository. http://github.com/ohdsi/achilles Accessed 2014-12-08.
[7] ACHILLES Demonstration. http://ohdsi.org/web/achilles/ Accessed 2014-12-08.
[8] OHDSI Forums http://forums.ohdsi.org Accessed 2014-12-08.
[9] Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, Suchard MA, DuMouchel W, Berlin JA. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol. 2013;178(4):645-651.
[10] Ryan PB, Madigan D, Stang PE, Schuemie MJ, Hripcsak G. Medication-wide association studies. CPT: Pharmacometrics & Systems Pharmacology 2013;2,e76;doi:10.1038/psp.2013.52.
[11] Longhurst CA, Harrington RA, Shah NH. A 'green button' for using aggregate patient data at the point of care. Health Aff (Millwood), 2014;33(7):1229-1235.
[12]Feinstein AR, Rubinstein JF, Ramshaw WA. Estimating prognosis with the aid of a conversational-mode computer program. Ann Intern Med. 1972;76(6):911-921.
[13]PCORnet: the national patient-centered clinical research network. http://www.pcornet.org, Accessed 2014-12-08.

Address for correspondence

George Hripcsak, MD, MS
Profesor and Chair of Biomedical Informatics
Columbia University Medical Center
622 W 168th St, PH20
New York, NY 10032
hripcsak@columbia.edu