Welcome to OHDSI 2019: This is our community

George Hripcsak, MD, MS, Chair of the Department of Biomedical Informatics at Columbia University Medical Center

Harlan M. Krumholz, MD, SM, Harold H. Hines, Jr. Professor of Medicine at Yale University School of Medicine; Director, Yale New Haven Hospital Center for Outcomes Research and Evaluation (CORE)
@hmkyale
How do we know we are making progress on our journey?
Thank you to our sponsors!

And contributions from viewers like you!

We thank the FDA for their generous support of the 2019 OHDSI symposium through the FDA SCIENTIFIC CONFERENCE GRANT PROGRAM (R13)
OHDSI is an open science community
To improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care.
• **Innovation**: Observational research is a field which will benefit greatly from disruptive thinking. We actively seek and encourage fresh methodological approaches in our work.

• **Reproducibility**: Accurate, reproducible, and well-calibrated evidence is necessary for health improvement.

• **Community**: Everyone is welcome to actively participate in OHDSI, whether you are a patient, a health professional, a researcher, or someone who simply believes in our cause.

• **Collaboration**: We work collectively to prioritize and address the real world needs of our community’s participants.

• **Openness**: We strive to make all our community’s proceeds open and publicly accessible, including the methods, tools and the evidence that we generate.

• **Beneficence**: We seek to protect the rights of individuals and organizations within our community at all times.
OHDSI community

We’re all in this journey together...

256 collaborators in 27 different countries over six continents
OHDSI’s community engagement

• Active community online discussion: forums.ohdsi.org
  – >2,770 distinct users have made >18,700 posts on >3,250 topics
  – Implementers, Developers, Researchers, CDM Builders, Vocabulary users, OHDSI in Korea, OHDSI in China, OHDSI in Europe

• Weekly community web conferences for all collaborators to share their research ideas and progress

• >25 workgroups for solving shared problems of interest
  – ex: Common Data Model, Population-level Estimation, Patient-level Prediction, Phenotype, NLP, GIS, Oncology, Women of OHDSI

• Quarterly tutorials in OHDSI tools and best practices, taught by OHDSI collaborators for OHDSI collaborators

• OHDSI Symposia held annually in North America, Europe and Asia to provide the community face-to-face opportunities to showcase research collaborations

• Follow us on Twitter @OHDSI and LinkedIn
In the last year, we’ve seen tremendous activity and interest across a wide range of topics in multiple categories (Implementers, Vocabulary Users, Developers, Researchers).
Diversity of the OHDSI community represented today at the OHDSI Symposium

- **Stakeholder group**
  - Academia
  - Government
  - Health System
  - Technology
  - Patient
  - Pharmaceutical
  - Payer

- **Disciplinary perspective**
  - Informatics
  - Computer Science
  - Medicine
  - Epidemiology
  - Statistics
  - Health Policy
  - Clinical Sciences

**Relationship with OHDSI community**

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am new to OHDSI and curious to learn more</td>
<td>240</td>
</tr>
<tr>
<td>I actively participate in OHDSI meetings and work groups</td>
<td>177</td>
</tr>
<tr>
<td>I use OHDSI tools and methods to support my research</td>
<td>176</td>
</tr>
<tr>
<td>I have an OMOP CDM instance</td>
<td>125</td>
</tr>
<tr>
<td>I am in the process of converting my data into the OMOP CDM</td>
<td>95</td>
</tr>
<tr>
<td>I actively participate in discussions on the OHDSI forum</td>
<td>74</td>
</tr>
<tr>
<td>I am participating in an OHDSI network research study</td>
<td>55</td>
</tr>
<tr>
<td>I contribute code to the OHDSI GitHub</td>
<td>48</td>
</tr>
</tbody>
</table>
OHDSI is
an international data network
Data across the OHDSI community

- 152 entries on [2019 OHDSI data network inventory](#)
- 133 different databases with patient-level data from various perspectives:
  - Electronic health records, administrative claims, hospital systems, clinical registries, health surveys, biobanks
- Data in 18 different countries, with >369 million patient records from outside US

All using one open community data standard: OMOP Common Data Model
Open community data standard: OMOP CDM v6

- Standardized clinical data
  - Person
    - Observation_period
    - Visit_occurrence
    - 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

- Standardized health economics
  - Cost
  - Payer_plan_period

- 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

- Results Schema
  - Cohort
  - Cohort_definition

- Concept relationship
  - Concept_ancestor
  - Concept_synonym
  - Concept_rela7onship

- Vocabulary
  - Standardized vocabularies

- Domain
  - Standardized vocabularies

- Concept_class
  - Standardized vocabularies

- Concept_relationship
  - Standardized vocabularies

- Relationship
  - Standardized vocabularies

- Concept_synonym
  - Standardized vocabularies

- Concept_ancestor
  - Standardized vocabularies

- Source_to_concept_map
  - Standardized vocabularies

- Drug_strength
  - Standardized vocabularies
OHDSI’s standardized vocabularies

• >130 Vocabularies across 40 domains
  – MU3 standards: SNOMED, RxNorm, LOINC
  – Disparate sources: ICD9CM, ICD10(CM), Read, NDC, Gemscript, CPT4, HCPCS…

• >7.4 million concepts
  – >3.0 million standard concepts
  – >3.8 million source codes
  – >511,000 classification concepts

• >45 million concept relationships
• >74 million ancestral relationships

Publicly available for download at: http://athena.ohdsi.org/
OHDSI is advancing science
What is OHDSI’s strategy to deliver reliable evidence?

• **Methodological research**
  – Develop new approaches to observational data analysis
  – Evaluate the performance of new and existing methods
  – Establish empirically-based scientific best practices

• **Open-source analytics development**
  – Design tools for data transformation and standardization
  – Implement statistical methods for large-scale analytics
  – Build interactive visualization for evidence exploration

• **Clinical evidence generation**
  – Identify clinically-relevant questions that require real-world evidence
  – Execute research studies by applying scientific best practices through open-source tools across the OHDSI international data network
  – Promote open-science strategies for transparent study design and evidence dissemination
Highlights of progress from the community: Data standards

- Increased adoption of OMOP CDM
- Evaluation of vocabulary
- Expanded vocabulary
- Community collaboration around conventions (THEMIS)
- Added rigor around data quality (see Clair and Andrew)
Research and Applications

Effect of vocabulary mapping for conditions on phenotype cohorts

George Hripcsak,1,2,3 Matthew E Levine,1,2 Ning Shang,1,2 and Patrick B Ryan1,2,4

1Department of Biomedical Informatics, Columbia University, New York, New York, USA, 2Observational Health Data Sciences and Informatics (OHDSI), New York, New York, USA, 3Medical Informatics Services, New York-Presbyterian Hospital, New York, New York, USA, and 4Epidemiology Analytics, Janssen Research and Development, Titusville, New Jersey, USA

Corresponding Author: George Hripcsak, MD, MS, Department of Biomedical Informatics, Columbia University Irving Medical Center, 622 W 168th St, PH38, New York, NY 10032, USA (hripcsak@ccolumbia.edu)

Received 27 April 2019, Revised 13 August 2019, Editorial Decision 22 August 2019, Accepted 3 September 2019

ABSTRACT

Objective: To study the effect on patient cohorts of mapping condition (diagnosis) codes from source billing vocabularies to a clinical vocabulary.

Materials and Methods: Nine International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9-CM) concept sets were extracted from eMERGE network phenotypes, translated to Systematized Nomenclature of Medicine - Clinical Terms concept sets, and applied to patient data that were mapped from source ICD9-CM and ICD10-CM codes to Systematized Nomenclature of Medicine - Clinical Terms codes using Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) vocabulary mappings. The original ICD9-CM concept set and a concept set extended to ICD10-CM were used to create patient cohorts that served as gold standards.

Results: Four phenotype concept sets were able to be translated to Systematized Nomenclature of Medicine - Clinical Terms without ambiguities and were able to perform perfectly with respect to the gold standards. The other 5 lost performance when 2 or more ICD9-CM or ICD10-CM codes mapped to the same Systematized Nomenclature of Medicine - Clinical Terms code. The patient cohorts had a total error (false positive and false negative) of up to 0.15% compared to querying ICD9-CM source data and up to 0.26% compared to querying ICD9-CM and ICD10-CM data. Knowledge engineering was required to produce that performance; simple automated methods to generate concept sets had errors up to 10% (one outlier at 25%).

Discussion: The translation of data from source vocabularies to Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) resulted in very small error rates that were an order of magnitude smaller than other error sources.

Conclusion: It appears possible to map diagnoses from disparate vocabularies to a single clinical vocabulary and carry out research using a single set of definitions, thus improving efficiency and transportability of research.
HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model

Jeremy L. Warner\textsuperscript{a,b,*}, Dmitry Dymshyts\textsuperscript{c}, Christian G. Reich\textsuperscript{d}, Michael J. Gurley\textsuperscript{e}, Harry Hochheiser\textsuperscript{f}, Zachary H. Moldwin\textsuperscript{g}, Rimma Belenkaya\textsuperscript{h}, Andrew E. Williams\textsuperscript{i}, Peter C. Yang\textsuperscript{h,j}

\textsuperscript{a}Vanderbilt University Medical Center, Nashville, TN, United States
\textsuperscript{b}HemOnc.org, LLC, Lexington, MA, United States
\textsuperscript{c}Odyssea Data Services, Inc., Cambridge, MA, United States
\textsuperscript{d}IQVIA, Cambridge, MA, United States
\textsuperscript{e}Northwestern University, Chicago, IL, United States
\textsuperscript{f}University of Pittsburgh, Pittsburgh, PA, United States
\textsuperscript{g}University of Illinois at Chicago College of Pharmacy, Chicago, IL, United States
\textsuperscript{h}Memorial Sloan Kettering Cancer Center, New York, NY, United States
\textsuperscript{i}Tufts University, Medford, MA, United States
\textsuperscript{j}Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States
Highlights of progress from the community: Methods research

• Phenotype definition
• Phenotype evaluation
• Study design evaluation
Facilitating phenotype transfer using a common data model

George Hripcsak\textsuperscript{a,b,*}, Ning Shang\textsuperscript{a}, Peggy L. Peissig\textsuperscript{c}, Luke V. Rasmussen\textsuperscript{d}, Cong Liu\textsuperscript{a}, Barbara Benoit\textsuperscript{e}, Robert J. Carroll\textsuperscript{f}, David S. Carrell\textsuperscript{g}, Joshua C. Denny\textsuperscript{h,*}, Ozan Dikilitas\textsuperscript{i}, Vivian S. Gainer\textsuperscript{e}, Kayla Marie Howell\textsuperscript{j}, Jeffrey G. Klann\textsuperscript{k}, Ifitkhar J. Kullo\textsuperscript{i}, Todd Lingren\textsuperscript{k}, Frank D. Mentch\textsuperscript{l}, Shawn N. Murphy\textsuperscript{m}, Karthik Natarajan\textsuperscript{a,b}, Jennifer A. Pacheco\textsuperscript{n}, Wei-Qi Wei\textsuperscript{f}, Ken Wiley\textsuperscript{m}, Chunhua Weng\textsuperscript{a}

\textsuperscript{a} Department of Biomedical Informatics, Columbia University, New York, NY, United States
\textsuperscript{b} Medical Informatics Services, NewYork-Presbyterian Hospital, New York, NY, United States
\textsuperscript{c} Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI, United States
\textsuperscript{d} Northwestern University Feinberg School of Medicine, Chicago, IL, United States
\textsuperscript{e} Research Information Science and Computing, Partners Healthcare, Boston, MA, United States
\textsuperscript{f} Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
\textsuperscript{g} Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States
\textsuperscript{h} Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
\textsuperscript{i} Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States
\textsuperscript{j} Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, United States
\textsuperscript{k} Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
\textsuperscript{l} Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA, United States
\textsuperscript{m} National Human Genome Research Institute, NIH, Bethesda, MD, United States
PheValuator: Development and evaluation of a phenotype algorithm evaluator

Joel N. Swerdel, George Hripcsak, Patrick B. Ryan

Background: The primary approach for defining disease in observational healthcare databases is to construct phenotype algorithms (PAs), rule-based heuristics predicated on the presence, absence, and temporal logic of clinical observations. However, a complete evaluation of PAs, i.e., determining sensitivity, specificity, and positive predictive value (PPV), is rarely performed. In this study, we propose a tool (PheValuator) to efficiently estimate a complete PA evaluation.

Methods: We used 4 administrative claims datasets: OptumInsight's de-identified Clininformatics "Datamart" (Eden Prairie, MN); IBM MarketScan Multi-State Medicaid; IBM MarketScan Medicare Supplemental Beneficiaries; and IBM MarketScan Commercial Claims and Encounters from 2000 to 2017. Using PheValuator involves (1) creating a diagnostic predictive model for the phenotype, (2) applying the model to a large set of randomly selected subjects, and (3) comparing each subject's predicted probability for the phenotype to inclusion/exclusion in PAs. We used the predictions as a "probabilistic gold standard" measure to classify positive/negative cases. We examined 4 phenotypes: myocardial infarction, cerebral infarction, chronic kidney disease, and atrial fibrillation. We examined several PAs for each phenotype including 1-time (1X) occurrence of the diagnosis code in the subject's record and 1-time occurrence of the diagnosis in an inpatient setting with the diagnosis code as the primary reason for admission (1X-IP-1stPos).

Results: Across phenotypes, the 1X PA showed the highest sensitivity/lowest PPV among all PAs. 1X-IP-1stPos yielded the highest PPV/lowest sensitivity. Specificity was very high across algorithms. We found similar results between algorithms across datasets.

Conclusion: PheValuator appears to show promise as a tool to estimate PA performance characteristics.
A plea to stop using the case-control design in retrospective database studies

Martijn J. Schuemie1,2,3 | Patrick B. Ryan1,4 | Kenneth K.C. Man5,6,7,8 | Ian C.K. Wong5,6 | Marc A. Suchard1,3,9,10 | George Hripcsak1,4,11

1Observational Health Data Sciences and Informatics, New York, New York
2Epidemiology Analytics, tuasen research and development, Titusville, New Jersey
3Department of Biostatistics, University of California, Los Angeles, California
4Department of Biomathematics, University of California, Los Angeles, California
5Department of Biostatistics, Columbia University Medical Center, New York, New York
6Centre for Safe Medication Practice and Research, Department of Pharmacy and Pharmacology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong
7research department of practice and Policy, UCL School of Pharmacy, London, UK
8Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
9Department of Social Work and Social Administration, Faculty of Social Science, The University of Hong Kong, Pokfulam, Hong Kong
10Department of Biostatistics, University of California, Los Angeles, California

The case-control design is widely used in retrospective database studies, often leading to spectacular findings. However, results of these studies often cannot be replicated, and the advantage of this design over others is questionable. To demonstrate the shortcomings of applications of this design, we replicate two published case-control studies. The first investigates isotretinoin and ulcerative colitis using a simple case-control design. The second focuses on dipeptidyl peptidase-4 inhibitors and acute pancreatitis, using a nested case-control design. We include large sets of negative control exposures (where the true odds ratio is believed to be 1) in both studies. Both replication studies produce effect size estimates consistent with the original studies, but also generate estimates for the negative control exposures showing substantial residual bias. In contrast, applying a self-controlled design to answer the same questions using the same data reveals far less bias. Although the case-control design in general is not at fault, its application in retrospective database studies, where all exposure and covariate data for the entire cohort are available, is unnecessary, as other alternatives such as cohort and self-controlled designs are available. Moreover, by focusing on cases and controls it opens the door to inappropriate comparisons between exposure groups, leading to confounding for which the design has few options to adjust for. We argue that this design should no longer be used in these types of data. At the very least, negative control exposures should be used to prove that the concerns raised here do not apply.
Highlights of progress from the community:
Open source development

- ATLAS 2.7.3 released
- Criteria2Query published
- Community contributions for multiple OMOP CDM utilities
Welcome to ATLAS.
ATLAS is an open source application developed as a part of OHDSI intended to provide a unified interface to patient level data and analytics.

Documentation
The ATLAS user guide can be found here.

Getting Started
- Define a New Cohort
- Search the Vocabulary

Begin performing research by defining the group of people you intend to study
Search the different ontologies used to describe patient level data around the world

Release Notes
ATLAS Version 2.7.3 Release Notes
WebAPI Version 2.7.3 Release Notes

This latest release contains 7 feature enhancements and issue resolutions:
- Cohort definitions creation data is 4 hours greater than actual while being on EST timezone
- Do not call user/refresh endpoint case of IAP authentication
- Characterization pop-up shows wrong percentage
- Role import / export works incorrectly
- Title Consistency
- Active Directory groups mapping issue
- Cannot save concept set modification in cohort definition

Apache 2.0
open source software
provided by
OHDSI
join the journey
Research and Applications

Criteria2Query: a natural language interface to clinical databases for cohort definition

Chi Yuan,1,2 Patrick B. Ryan,1,2 Casey Ta,1 Yixuan Guo,1 Ziran Li,1 Jill Hardin,2 Rupa Makadia,2 Peng Jin,1 Ning Shang,1 Tian Kang,1 and Chunhua Weng1

1Department of Biomedical Informatics, Columbia University, New York, New York, USA, 2Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, Jiangsu Province, P.R. China, and 3Epidemiology Analytics, Janssen Research and Development, Titusville, New Jersey, USA

Corresponding Author: Chunhua Weng, PhD, Department of Biomedical Informatics, Columbia University, 622 West 168th Street, PH-29, Room 497, New York, NY 10032, USA (chunhua@columbia.edu)

Received 7 September 2018; Revised 16 November 2018; Editorial Decision 27 November 2018; Accepted 29 November 2018

ABSTRACT

Objective: Cohort definition is a bottleneck for conducting clinical research and depends on subjective decisions by domain experts. Data-driven cohort definition is appealing but requires substantial knowledge of terminologies and clinical data models. Criteria2Query is a natural language interface that facilitates human-computer collaboration for cohort definition and execution using clinical databases.

Materials and Methods: Criteria2Query uses a hybrid information extraction pipeline combining machine learning and rule-based methods to systematically parse eligibility criteria text, transforms it first into a structured criteria representation and next into sharable and executable clinical data queries represented as SOL queries conforming to the OMOP Common Data Model. Users can interactively review, refine, and execute queries in the ATLAS web application. To test effectiveness, we evaluated 125 criteria across different disease domains from ClinicalTrials.gov and 52 user-entered criteria. We evaluated F1 score and accuracy against 2 domain experts and calculated the average computation time for fully automated query formulation. We conducted an anonymous survey evaluating usability.

Results: Criteria2Query achieved 0.795 and 0.805 F1 score for entity recognition and relation extraction, respectively. Accuracies for negation detection, logic detection, entity normalization, and attribute normalization were 0.984, 0.964, 0.914, and 0.793, respectively. Fully automatic query formulation took 1.22 seconds/criterion. More than 90% (11 of 13) of users would use Criteria2Query in their future cohort definition tasks.

Conclusions: We contribute a novel natural language interface to clinical databases. It is open source and supports fully automated and interactive modes for autonomous data-driven cohort definition by researchers with minimal human effort. We demonstrate its promising user friendliness and usability.
Data and text mining

PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model

Benjamin S. Glicksberg 1, Boris Oskotsky 1, Phyllis M. Thangaraj 2,3,4,†, Nicholas Giangreco 2,3,4,†, Marcus A. Badgeley 5,†, Kipp W. Johnson 5,†, Debajyoti Datta 1, Vivek A. Rudrapatna 1,6, Nadav Rappoport 1, Mark M.Shervey 5, Riccardo Miotto 5, Theodore C. Goldstein 1, Eugenia Rutenberg 1, Remi Fraizer 7, Nelson Lee 2, Sharat Israni 1, Rick Larsen 7, Bethany Percha 7, Li Li 5, Joel T. Dudley 5, Nicholas P. Tatonetti 2,3,4,† and Atul J. Butte 1,8,*

1Baker Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA 94158, USA, 2Department of Biomedical Informatics, 3Department of Systems Biology, 4Department of Medicine, Columbia University, New York, NY 10032, USA, 5Departments of Genomics and Data Science, Icahn Institute for Genomic Sciences and Multiscale Biology, Icahn School of Medicine at Mount Sinai, Institute of Next Generation Healthcare, New York, NY 10029, USA, 6Division of Gastroenterology, Department of Medicine, University of California, San Francisco, CA 94158, USA, 7Enterprise Informatics and Analytics, University of California, San Francisco, San Francisco, CA 94158, USA and 8Center for Data-Driven Insights and Innovation, University of California Health, Oakland, CA 94607, USA

Application Notes

ROMOP: a light-weight R package for interfacing with OMOP-formatted electronic health record data

Benjamin S. Glicksberg,† Boris Oskotsky, Nicholas Giangreco,† Phyllis M. Thangaraj, Vivek Rudrapatna,† Debajyoti Datta, Remi Fraizer, Nelson Lee, Rick Larsen, Nicholas P. Tatonetti† and Atul J. Butte†

1Department of Pediatrics, Baker Computational Health Sciences Institute, University of California San Francisco, San Francisco, California, USA, 2Department of Biomedical Informatics, Systems Biology, and Medicine, Columbia University, New York, New York, USA and 3Academic Research Systems, Department of Enterprise Data Warehouse University of California San Francisco, San Francisco, California, USA

†These two authors contributed equally to the study.

Corresponding Author: Atul J. Butte, MD, PhD, Baker Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA 94158, USA (Atul.Butte@ucsf.edu)

Received 3 July 2018, Revised 26 October 2018, Editorial Decision 25 November 2018, Accepted 2 December 2018
Highlights of progress from the community: Clinical applications

LEGEND workgroup focus on clinically important hypotheses:
- ACE vs. THZ
- ACE vs. ARB
- Beta blocker vs. first-line
- Chlorthalidone vs. hydrochlorothiazide
- Mono vs. combo therapy

http://data.ohdsi.org/LegendBasicViewer/
OHDSI is building collaborations
FDA Biologics Effectiveness and Safety (BEST) Initiative

Biologics Effectiveness and Safety (BEST) Initiative: Incorporating ISBT-128 Codes into OHDSI's OMOP Common Data Model to Build a National Hemovigilance System to Monitor Transfusion-Related Adverse Events

Joyce Olid1, Kinnara Chada1, Joanne Gruber1, Graça Dore2, Alan Williams3, Emily Storch3, Juan M Banda4, Sarath Gombar3, Deepa Balraj1, Ross Hayden1, Paul Blodgett1, Shaun Grinnell1, George Hripcsak3, Thomas Falcozner1, Karthik Katakanji5, Dmitry Dymshity5, Sara Dempster3, Christian Reich2, Nandini Selvam7, Nerissa Williams6, Steven Anderson1, Azadeh Shoab1

1Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA; 2Stanford University, Stanford, CA, USA; 3Regenstrief Institute, Indianapolis, Indiana, USA; 4Columbia University, New York, NY, USA; 5Observational Health Data Sciences and Informatics, New York, NY, USA; 6Plexxys Data Services Inc., Cambridge, MA, USA; 7IQVIA, Cambridge, MA, USA.

INTRODUCTION

The U.S. FDA Center for Biologics Evaluation and Research (CBER) regulates collection of whole blood and blood components utilized in transfusion.

OBJECTIVE

The aim of this study was to build a component of the infrastructure for a national hemovigilance system using EHR data sources to monitor transfusion-related AEs by incorporating the ISBT-128 coding system into the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) of the Observational Health Data Sciences and Informatics (OHDSI) consortium.

METHODS

The CBER BEST Initiative is a collaboration with IQVIA, OHDSI Consortium, Columbia University, Stanford University, Indiana University, Regenstrief Institute, Georgia Institute of Technology, and University of California Los Angeles. Within the BEST Initiative, we used three EHR databases that cover approximately 24 million patient records from geographically diverse areas of the U.S. We added a library of 14,543 ISBT-128 codes to the OMOP CDM. Each EHR data source requested access to its corresponding blood bank data and transformed its data into the OMOP CDM containing the newly added ISBT-128 codes. By querying the databases, we determined the type and frequency of ISBT-128 codes used in patient records from 2010-2017 within the blood banks of EHR data providers participating in the BEST Initiative.
NIH All of Us Research Program

• 1,000,000 diverse participants
• Clinical data in OMOP CDM

The future of health begins with you

The All of Us Research Program is a historic effort to gather data from one million or more people 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.

JOIN NOW
RESEARCH ARTICLE

Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model

Jeffrey G. Klann 1,2,3, Matthew A. H. Joss 1, Kevin Embree 4, Shawn N. Murphy 1,2,3

1 Research Information Science and Computing, Partners Healthcare, Boston, Massachusetts, United States of America, 2 Harvard Medical School, Boston, Massachusetts, United States of America, 3 Laboratory of Computer Science, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America, 4 Personalized Medicine, Partners Healthcare, Boston, Massachusetts, United States of America

* jeff.klann@mgh.harvard.edu

Abstract

Background

The All Of Us Research Program (AOU) is building a nationwide cohort of one million patients’ EHR and genomic data. Data interoperability is paramount to the program’s success. AOU is standardizing its EHR data around the Observational Medical Outcomes Partnership (OMOP) data model. OMOP is one of several standard data models presently used in national-scale initiatives. Each model is unique enough to make interoperability difficult. The i2b2 data warehouse and analytics platform is used over 200 sites worldwide, which uses a flexible ontology-driven approach for data storage. We previously demonstrated this ontology system can drive data reconfiguration, to transform data into new formats without site-specific programming. We previously implemented this on our 12-site Accessible Research Commons for Health (ARCH) network to transform i2b2 into the Patient Centered Outcomes Research Network model.
Electronic Medical Records and Genomics (eMERGE) Network
Facilitating phenotype transfer using a common data model

George Hripcsak\textsuperscript{a,b,}\textsuperscript{*}, Ning Shang\textsuperscript{a}, Peggy L. Peis\textsuperscript{a}, Barbara Benoit\textsuperscript{a}, Robert J. Carroll\textsuperscript{b}, David S. Califano\textsuperscript{b}, Vivian S. Gainer\textsuperscript{e}, Kayla Marie Howell\textsuperscript{f}, Jeffrey Frank D. Mentch\textsuperscript{f}, Shawn N. Murphy\textsuperscript{e}, Karthik\textsuperscript{f}, Ken Wiley\textsuperscript{m}, Chunhua Weng\textsuperscript{a}

\textsuperscript{a} Department of Biomedical Informatics, Columbia University, New York, NY, United States
\textsuperscript{b} Medical Informatics Services, NewYork-Presbyterian Hospital, New York, NY, United States
\textsuperscript{c} Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marschfield, WI, United States
\textsuperscript{d} Northwestern University Feinberg School of Medicine, Chicago, IL, United States
\textsuperscript{e} Research Information Science and Computing, Partners Healthcare, Boston, MA, United States
\textsuperscript{f} Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
\textsuperscript{g} Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States
\textsuperscript{h} Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
\textsuperscript{i} Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States
\textsuperscript{j} Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University
\textsuperscript{k} Cincinnati Children’s Hospital Medical Center, Cincinnati, OH, United States
\textsuperscript{l} Center for Applied Genomics, Children’s Hospital of Philadelphia, Philadelphia, PA, United States
\textsuperscript{m} National Human Genome Research Institute, NIH, Bethesda, MD, United States

\textbf{Background:} Implementing clinical phenotypes across a network is labor intensive and potentially error prone. Use of a common data model may facilitate the process.

\textbf{Methods:} Electronic Medical Records and Genomics (eMERGE) sites implemented the Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) Common Data Model across their electronic health record (EHR)-linked DNA biobanks. Two previously implemented eMERGE phenotypes were converted to OMOP and implemented across the network.

\textbf{Results:} It was feasible to implement the common data model across sites, with laboratory data producing the greatest challenge due to local encoding. Sites were then able to execute the OMOP phenotype in less than one day, as opposed to weeks of effort to manually implement an eMERGE phenotype in their bespoke research EHR databases. Of the sites that could compare the current OMOP phenotype implementation with the original eMERGE phenotype implementation, specific agreement ranged from 100\% to 43\%, with disagreements due to the original phenotype, the OMOP phenotype, changes in data, and issues in the databases. Using the OMOP query as a standard comparison revealed differences in the original implementations despite starting from the same definitions, code lists, flowcharts, and pseudocode.

\textbf{Conclusion:} Using a common data model can dramatically speed phenotype implementation at the cost of having to populate that data model, though this will produce a net benefit as the number of phenotype implementations increases. Inconsistencies among the implementations of the original queries point to a potential benefit of using a common data model so that actual phenotype code and logic can be shared, mitigating human error in re-interpretation of a narrative phenotype definition.
The European Health Data and Evidence Network

**Vision**

The European Health Data & Evidence Network (EHDEN) aspires to be the trusted observational research ecosystem to enable better health decisions, outcomes and care.

**Mission**

Our mission is to provide a new paradigm for the discovery and analysis of health data in Europe, by building a large-scale, federated network of data sources standardised to a common data model.
Objectives

Harmonisation
Harmonise in excess of 100 million anonymised health records to the OMOP common data model, supported by an ecosystem of certified SMEs, and technical architecture for a federated network.

Evidence
Impact our understanding of, and improvement of, clinical outcomes for patients within diverse healthcare systems in the EU.

Community
Establish a self-sustaining open science collaboration in Europe, supporting academia, industry, regulators, payers, government, NGOs and others.
National CDM Projects in Korea
2018-2022

- Clinical Trial support system
- AI service for Diagnosis and Prescription
- CDM rules & quality improvement
  - 2 projects, 3 years, $500K each
- CDM standardization & de-identification guidelines
  - 1 project, 4 years, $1.5M
- Public Purpose Researches on Health policy / Health Technology
  - 10 projects, 3 years, $500K each
- Privacy Protection / Law
  - 10 projects, 3 years, $3500K each

Projects for Industry Services
(Ministry of Industry)
6 projects, 3 years, $1.5M for each

Projects for CDM standard/policy
(Ministry of Industry / Health & Welfare)

FEEDER-NET Platform

- Tertiary hospital-centered data network
  - 1st project, 2018-2020
    - 40 hospitals, 54M patients
- Tertiary Teaching hospitals
- General hospitals
- OMOP CDM
- OMOP CDM
- OMOP CDM

FEEDER-NET: FEderated E-health big Data for Evidence Renovation NETwork
FEEDER-NET Data Network in Korea

Data Network of 60+ Hospitals, 98M Patients

70% of Tertiary Teaching Hospitals
OHDSI is

a community of collaborators
5 day event
~30 researchers
Result: 2 papers

Results of the study available at:
http://data.ohdsi.org/oxfordMortalityExternalValidation/

External Validation package available at:
https://github.com/OHDSI/
StudyProtocolSandbox/tree/master/
mortalityValidation
Case Western Reserve University:
OHDSI face-to-face documentation-a-thon
OHDSI China Symposium 2019
- A platform to stimulate community building: 250 participants from 27 countries
- OHDSI Europe in action: 35 posters, 8 software demos
- Educate and train the community: 5 full day tutorials

www.ohdsi-europe.org
Fudan University – OHDSI tutorials
OHDSI Korea – Study design datathon
Upcoming symposia
OHDSI Korea Symposium

12-14th December 2019,
KONJIAM Resort, Gwangju, Gyeonggi-Do, Republic of Korea
European OHDSI Symposium 2020

27-29 March 2020 – Oxford, UK

Mathematical Institute, University of Oxford
2019 Theme: Continuous evaluation

How do we know we are making progress on our journey?
OHDSI evaluates itself and publishes the results

- OMOP CDM vocabulary evaluation
  - Automated translation of database works
  - Best not to automated the translation of cohort definitions

- eMERGE phenotype implementation
  - Without CDM, narrative+flowchart+pseudocode+code list -> inconsistent
  - With CDM, can improve consistency and efficiency but caveats

- PheValuator phenotype evaluation
  - Can estimate performance without manually curating gold standard
  - Estimates are imperfect
Today’s focus

How do we know we are making progress toward using real-world evidence for regulatory decision-making?
Symposium • Day 2

<table>
<thead>
<tr>
<th>Time</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>7:30 - 8:00 am</td>
<td>Registration with light breakfast (Grand Ballroom F-H Foyer)</td>
</tr>
<tr>
<td>8:00 - 9:00 am</td>
<td>Welcome to OHDSI 2019: This is our community (Grand Ballroom DE)</td>
</tr>
<tr>
<td></td>
<td>• George Hripcsak, MD, MS, Vivian Beaumont Allen Professor and Chair of Biomedical Informatics at Columbia University Irving Medical Center; Director of Medical Informatics Services at NewYork Presbyterian Hospital/Columbia Campus</td>
</tr>
<tr>
<td></td>
<td>• Harlan Krumholz, MD, Harold H. Hines, Jr. Professor of Medicine, Epidemiology and Public Health at Yale University</td>
</tr>
<tr>
<td>9:00 - 11:00 am</td>
<td>Plenary Session: A journey toward real-world evidence for regulatory decision-making Building confidence in real-world data: Data quality reporting</td>
</tr>
<tr>
<td></td>
<td>• Clair Blacketer, MPH, PMP, Associate Director of Epidemiology Analytics at Janssen Research &amp; Development; PhD Student at Erasmus Medical Center Rotterdam</td>
</tr>
<tr>
<td></td>
<td>• Andrew Williams, PhD, Senior Informatics Advisor at Tufts Medical Center</td>
</tr>
<tr>
<td></td>
<td>Establishing scientific best practices for real-world analysis: Book Of OHDSI</td>
</tr>
<tr>
<td></td>
<td>• Martijn Schuemie, PhD, Senior Director and Research Fellow of Epidemiology Analytics at Janssen Research &amp; Development; Visiting Scholar of Biostatistics at the University of California, Los Angeles</td>
</tr>
<tr>
<td></td>
<td>• David Madigan, PhD, Professor of Statistics at Columbia University</td>
</tr>
<tr>
<td></td>
<td>Proving reliable real-world evidence: Replicating RCTs using LEGEND</td>
</tr>
<tr>
<td></td>
<td>• Patrick Ryan, PhD, Vice President of Observational Health Data Analytics at Janssen Research &amp; Development; Adjunct Assistant Professor of Biomedical Informatics at Columbia University</td>
</tr>
<tr>
<td></td>
<td>• George Hripcsak, MD, MS, Vivian Beaumont Allen Professor and Chair of Biomedical Informatics at Columbia University Irving Medical Center; Director of Medical Informatics Services at NewYork Presbyterian Hospital/Columbia Campus</td>
</tr>
<tr>
<td>11:00 - 11:30 am</td>
<td>Stakeholder panel: What has been done? Where should we go? How do we get there?</td>
</tr>
<tr>
<td></td>
<td>Moderator: Harlan M. Krumholz, MD, SM, Harold H. Hines, Jr. Professor of Medicine, Epidemiology and Public Health at Yale University</td>
</tr>
<tr>
<td></td>
<td>Panelists: Joseph Ross, MD, MHS, Professor of Internal Medicine at Yale University</td>
</tr>
<tr>
<td></td>
<td>Azadeh Shoaii, PhD, MHS, Lead of CBER Sentinel Program at US Food and Drug Administration</td>
</tr>
<tr>
<td></td>
<td>Fatemah Alnafal, MIT, Research Specialist at the Saudi Food and Drug Authority</td>
</tr>
<tr>
<td>11:30 - 1:00 pm</td>
<td>OHDSI Collaborator Showcase: Part 1 (Grand Ballroom F-H) Software demonstrations and poster presentations highlighting the scientific progress throughout the OHDSI community</td>
</tr>
</tbody>
</table>
A journey toward real-world evidence for real-world decision-making
The current medical research enterprise cannot keep pace with the information needs of patients, clinicians, administrators, and policy makers. The flow of new knowledge is too slow, and its scope is too narrow.
The medical research community’s delay in adopting big data approaches has left it particularly ill prepared for a precision medicine future that is designed to provide personalized information and individualized care.

Medicine aspires to a learning health care system, but is failing to rapidly learn from the data being generated through the course of clinical care.
Growing access to diverse 'real-world' data sources is enabling new approaches to close persistent evidence gaps about the optimal use of medical products in real-world practice.
Here, we argue that contrary to widespread impressions, existing FDA regulations embody sufficient flexibility to accommodate the emerging tools and methods needed to achieve this goal.
...Congressional mandate requires “the coordination of relevant Federal health programs to build data capacity for comparative clinical effectiveness research . . . from multiple sources, including electronic health records”

... governmental agencies and partners in the private sector, including those that fund research, are now collaborating on the focused development of infrastructure for the generation of evidence that can support a learning health system.
If RWD and RWE are to be effectively leveraged for public health purposes, there will need to be shared learning and collaboration across clinicians, patients, health care systems, pharmaceutical companies, and regulators.
What do I love about OHDSI?

• Spirit of collaboration, kindness, generosity
What do I love about OHDSI?

• Principles of transparency, open science, integrity
What do I love about OHDSI?

• Scientific rigor; reproducibility, validity
Key Challenges

- Data quality
- Data spectrum
- Causal inference
- Communication/Education
- Application
Next Session

• Data quality
  – Plenary #1: Blacketer and Williams

• Analytic standards
  – Plenary #2: Schuemie and Madigan

• Evidence quality
  – Plenary #3: Ryan and Hripcsak