April Olympians #3 / Tools to Evaluate ETL

OHDSI Community Call
April 16, 2024 • 11 am ET
## Upcoming Community Calls

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 16</td>
<td>April Olympians Update</td>
</tr>
<tr>
<td>April 23</td>
<td>April Olympians Update</td>
</tr>
<tr>
<td>April 30</td>
<td>April Olympians Update</td>
</tr>
<tr>
<td>May 7</td>
<td>DevCon 2024 Review</td>
</tr>
<tr>
<td>May 14</td>
<td>10-Minute Tutorials</td>
</tr>
<tr>
<td>May 21</td>
<td>Open Studies in the OHDSI Community</td>
</tr>
<tr>
<td>May 28</td>
<td>Collaborator Showcase Brainstorm</td>
</tr>
<tr>
<td>June 4</td>
<td><strong>NO CALL – EUROPEAN SYMPOSIUM</strong></td>
</tr>
<tr>
<td>June 11</td>
<td>European Symposium Review</td>
</tr>
<tr>
<td>June 18</td>
<td>Application of LLMs in Evidence Generation Process</td>
</tr>
<tr>
<td>June 25</td>
<td>Recent OHDSI Publications</td>
</tr>
</tbody>
</table>
Three Stages of The Journey

Where Have We Been?
Where Are We Now?
Where Are We Going?
OHDSI Shoutouts!

Congratulations to the team of Nhung TH Trinh, Annika M Jödicke, Martí Català, Núria Mercadé-Besora, Saeed Hayati, Angela Lupattelli, Daniel Prieto-Alhambra, and Hedvig ME Nordeng on the publication of "Effectiveness of COVID-19 vaccines to prevent long COVID: data from Norway in The Lancet Respiratory Medicine."
Congratulations to the team of Guy Tsafnat, Rachel Dunscombe, Davera Gabriel, Grahame Grieve, and Christian Reich on the publication of *Converge or Collide? Making Sense of a Plethora of Open Data Standards in Health Care* in the *Journal of Medical Internet Research*.
Three Stages of The Journey

Where Have We Been?
Where Are We Now?
Where Are We Going?
### Upcoming Workgroup Calls

<table>
<thead>
<tr>
<th>Date</th>
<th>Time (ET)</th>
<th>Meeting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday</td>
<td>12 pm</td>
<td>Common Data Model Vocabulary Subgroup</td>
</tr>
<tr>
<td>Tuesday</td>
<td>1 pm</td>
<td>Common Data Model</td>
</tr>
<tr>
<td>Wednesday</td>
<td>7 am</td>
<td>Medical Imaging</td>
</tr>
<tr>
<td>Wednesday</td>
<td>3 pm</td>
<td>Joint Vulcan/OHDS Meeting</td>
</tr>
<tr>
<td>Thursday</td>
<td>8 am</td>
<td>OHDSI India Community Call</td>
</tr>
<tr>
<td>Thursday</td>
<td>9 am</td>
<td>OMOP CDM Oncology Vocabulary/Development Subgroup</td>
</tr>
<tr>
<td>Thursday</td>
<td>9:30 am</td>
<td>Themis</td>
</tr>
<tr>
<td>Thursday</td>
<td>12 pm</td>
<td>HADES</td>
</tr>
<tr>
<td>Thursday</td>
<td>7 pm</td>
<td>Dentistry</td>
</tr>
<tr>
<td>Friday</td>
<td>10 am</td>
<td>GIS-Geographic Information System</td>
</tr>
<tr>
<td>Friday</td>
<td>10:30 am</td>
<td>Open-Source Community</td>
</tr>
<tr>
<td>Friday</td>
<td>11:30 am</td>
<td>Steering Group</td>
</tr>
<tr>
<td>Monday</td>
<td>9 am</td>
<td>Vaccine Vocabulary</td>
</tr>
<tr>
<td>Monday</td>
<td>4 pm</td>
<td>Eyecare &amp; Vision Research</td>
</tr>
<tr>
<td>Tuesday</td>
<td>9 am</td>
<td>OMOP CDM Oncology Genomic Subgroup</td>
</tr>
</tbody>
</table>
Maternal Health Data Science Fellowship

This program is designed to empower clinical investigators to leverage emerging technologies for improved maternal and neonatal care while reducing morbidity and mortality.

Three main components of this program:

1) **Career Development** (create evidence, leverage data models, build skills on network studies)

2) **Practice** (design effective observational research protocols, master tools, write papers/grants)

3) **Networking** (build relationships with mentors, learners, coordinate with global OHDSI collaborators)

Application deadline: May 15

Want to build your career?
Generate reproducible evidence by leading multi-institutional studies!

Learn more & apply!

@OHDSI www.ohdsi.org #JoinTheJourney
Next CBER BEST Seminar: Apr. 17

2021 Titan Award honoree Yong Chen will lead the next CBER BEST Seminar on Wednesday, April 17 (11 am-12 pm).

**Topic:** Real-World Effectiveness of BNT162b2 Against Infection and Severe Diseases in Children and Adolescents: causal inference under misclassification in treatment status.

Next CBER BEST Seminar: Apr. 17

CBER BEST Seminar Series

The CBEST Initiative Seminar Series is designed to share and discuss recent research of relevance to ongoing and future surveillance activities of CBERS. The seminars will provide information on characteristics of biologicals, required infrastructure, study designs, and analytic methods utilized for pharmacovigilance and pharmacoeconomic studies of biologicals. They will also cover information regarding potential data sources, informatics challenges and requirements, the utilization of real-world data and evidence, and risk-benefit analysis for biologic products. The length of each session may vary, and the presenters will be invited by outside FDA.

Below you will find details of upcoming CBERS seminars, including virtual links that will be open to anybody who wishes to attend. Speakers who give their consent to be recorded will also have their presentations included on this page; you can find those sessions below the list of upcoming speakers.

Upcoming Seminars

- April 17 (11 am ET): Yong Chen, University of Pennsylvania

Previous Seminars

- Jan. 17, 2024 - Anna Ostropoleos, Olympos Data Services
- Dec. 6, 2023 - Jenny Sun, Pfizer
- June 14, 2023 - Katsuyuki Bykow, Harvard Medical School
- May 3, 2023 - Xinlong Li and Daniel Prieto-Alhambra, University of Oxford, NIDCR and NIDRR
- Apr. 12, 2023 - Kristin Boland, P-90
- Mar. 22, 2023 - Marthe Schumie, Janssen R&D
- Feb. 8, 2023 - Fan Du, UCLA

ohdsi.org/cber-best-seminar-series

www.ohdsi.org  #JoinTheJourney
DevCon 2024: April 26, 9 am-3 pm ET

Morning Agenda

9:00 am – Introduction

9:15 am – Developers Panel and Lightning Talks (Katy Sadowski)
  • OHDSI/OMOP – The hard way is the easy way! (Prof. Vishnu V Chandrabalan)
  • Moving OMOP to the Cloud With DBT and Snowflake (Roger Carlson)
  • Use cases for ORMs in OMOP (Dr. Georgina Kennnedy)
  • Carrot: code-free OMOP ETL without full data access (Dr. Sam Cox)

10:45 am – Darwin EU® Developers Update (Adam Black)

12:00 pm – Break

Afternoon Agenda

12:30 pm – OHDSI Ecosystem Updates
  • TAB Update (Frank DeFalco)
  • Strategus Update (Anthony Sena)
  • Broadsea Update (Lee Evans)
  • Kheiron Updates (Paul Nagy)

1:15 pm – JACKALOPE PLUS The Power of ML for Healthcare Data Mapping & Management (Denys Kaduk)

2:00 pm - An Introduction to Knowledge Graphs using PheKnowLator and OMOP2OBO with Example Applications in Drug Surveillance and Computational Phenotyping (Tiffany Callahan)
The 2024 OHDSI Global Symposium will be held Oct. 22-24 at the Hyatt Regency Hotel in New Brunswick, NJ.

Tentative symposium format:
- **Oct. 22** – tutorials
- **Oct. 23** – plenaries, collaborator showcase
- **Oct. 24** – workgroup activities
Registration is now OPEN for the **2024 OHDSI Europe Symposium**, which will be held June 1-3 in Rotterdam, Netherlands.

- **June 1** – tutorial/workshop
- **June 2** – tutorial/workshop
- **June 3** – main conference

[ohdsi-europe.org]
Augmenting the National COVID Cohort Collaborative (N3C) Dataset with Medicare and Medicaid (CMS) Data, Secure and Deidentified Clinical Dataset

**INTRO:**
The National COVID Cohort Collaborative (N3C) data ecosystem is a platform that provides researchers access to COVID-related patient EHR data in OMOP CDM format. It is the largest centralized repository of COVID-related Patient EHR data in U.S. It is the largest OMOP instance to our knowledge. CMS claims data is also transformed into OMOP CDM format using code map terminology translation. N3C COVID patient cohort is now linked to CMS claims data via Privacy Preserving Record Linkage (PPLR). As a result, N3C EHR datasets in OMOP CDM format are enriched with the following additional CMS claims data.

**Input:**
- Inpatient
- Outpatient
- Part D drug prescription
- Part B
- Long term care
- Durable medical equipment
- Home health
- Skilled nursing
- Long Term Care

**Methods:**
1. CMS claim files in wide format are parsed and pivoted into long format. The clinical concept codes are organized into a condensed format per patient visit for efficient data transformation.
2. The condensed dataset is then used by the Code Map service to generate the clinical concept translation table. The unified version of the OMOP vocabulary tables are used to perform the translation from the source code to OMOP concept IDs.
3. The generated code map service table is used as input in the data pipeline to transform the CMS claims datasets into OMOP CDM format.
4. The data pipeline is built to generate CMS dataset in OMOP CDM format with N3C PPLR linkage.
5. N3C data is enriched with CMS data per PPLR-linked N3C patient. In cases where N3C person_id is duplicated, a Global ID is provided for each.

**Workflow to generate Combined OMOP from EHR & CMS data**

**Putting the patient back together again**
A timeline with no gaps, either overlapping or contiguous, is used to construct a "Macro Visit" akin to the N3C macro visit approach.

**RESULT:**
- # of patients in N3C: 20,864,732
- # of CMS patient claims: 213,096,774
- Total number of claims in N3C: 28,842,000
- Total number of claims in CMS: 20,864,732
- N3C dataset enriched by CMS data

**Terminology Mapping:**
Terminology codes appear in multiple columns, e.g., code1 to code4. And some claim source files were over 6,000 columns wide. The dataset is pivoted to coarsen format to provide the clinical concept translation table using OMOP vocabulary tables.
The Feasibility of Clinical Quality Language (CQL) Based Digital Quality Measures (dQMs) Implementation to OMOP CDM

Emir Amaro Syailendra, Woo Yeon Park, Ben Hamlin, Paul Nagy

The feasibility of using CQL-based digital quality measures (dQMs) to implement digital quality measures (dQMs) to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is evaluated. The CQL measures are converted to OMOP CDM using Atlas, a tool that translates CQL to OMOP CDM. The study demonstrates the feasibility of implementing dQMs to OMOP CDM and highlights the importance of standardization in healthcare data.
WEDNESDAY

Using Cohort Diagnostics to Assess the Phenotypic Data Quality in All of Us Research Program

(Lina Sulieman, Karthik Natarajan)

Using Cohort Diagnostics to Assess the Phenotypic Data Quality in All of Us Research Program

**INTRODUCTION**
- Quality of clinical data for research:
  - Inform the utility of the data
  - Affect reproducibility of clinical research
- Current quality metrics: general, not phenotype-specific
- Objective: Utilizing OHDSI Cohort Diagnostics to assess the phenotypic data quality in the All of Us Research Program, focusing on breast cancer

**METHODS**
- Applying phenotype library on All of Us Research Program, March 2022 release:
  - Identified two breast cancer cohorts:
    - Cohort 1: 41,285,981 patients
    - Cohort 2: 41,285,982
- Overlap between cohorts:
- Extracted the incident rates, time distributions, and covariates
- External validation: Compared temporal trends in breast cancer cohorts in all of Us Research Program to multiple databases

**RESULTS**
- All of Us dataset included 331,382 participants:
  - White, 35.83% participants
  - Female: 60% participants
- OHDSI breast cancer cohorts:
  - Cohort 1: 1995 participants
  - Cohort 2: 1995 participants
  - No patients younger than 39 years old (Fig 1)
- Median days before/after index diagnosis of breast cancer: All of Us higher than other datasets (Fig 2)
- Temporal trends (Table 1)
- Height, weight increased by 0.13 in “Stable” to 2071
- Hemoglobin, neuroticism doubled in “Mature” to 2071

**DISCUSSION**
- Cohort diagnostic: assessing the quality of phenotypic data in the research databases
- Observed data quality in breast cancer cohorts extracted from the All of Us Research Program:
  - Expected trends in age groups
  - Higher than expected incident rates in patients identified as other than female
  - Sex or Gender in All of Us different outcome, cautious when implementing phenotype library
  - Expected trends in temporal characteristics for breast cancer diagnosis
- Weight, height, hemoglobin, neuroticism: more labs were ordered after diagnosis. Measurements are taken periodically to assess the patient’s progress and prepare for treatment
- Drugs to treat breast cancer usage, much higher than before diagnosis

**Acknowledgment**
- All of Us Research Program, Participants
- All of Us Cohort Data and Research Centers

@OHDSI www.ohdsi.org

#OHDSISocialShowcase This Week

#JoinTheJourney
#OHDSISocialShowcase This Week

**THURSDAY**

Demonstration of the OHDSI phenotype library

*(Gowtham Rao)*

**Demonstration of the OHDSI phenotype library**

*PRESENTED: Gowtham Rao*

**Who Cares?**

Researchers: No more reinventing the wheel. Use peer-reviewed cohort definitions and get to your results faster.


Data Scientists: Garbage in, garbage out. Use quality definitions to elevate your findings.

Funders: Maximize ROI. Your grants power more reusable, standardized research assets.

**Bottom Line:** It’s more than a library; it’s a healthcare research game-changer. Community-built, peer-reviewed, and FAIR-compliant.

**How We Keep It Real**

Collected: Sourced from OHDSI forums. Email submissions to Rao. Attested: You say it’s good; we take your word (initially).

Tested: Run through PhenotypeLibrary/Diagnostics on OHDSI network.

**Quality Check**: Automated + human oversight.

Peer Review: Open forum or workshop discussions. Public scrutiny.

**Infrastructure**: Hosted on GitHub; R package. API-ready.

**Versioning**: DOI for every release. Cite us.

---

**Main Finding 1: FAIR*-Compliant Cohort Definition Repository**

**What:** The OHDSI Phenotype Library (PL) is an open, version-controlled repository designed to guide real-world evidence towards the FAIR principles—Findability, Accessibility, Reproducibility, and Interoperability.

**So What:** This compliance ensures that each cohort definition is easily searchable, version-controlled, and standardized for use across different studies.

**Now What:** Researchers can leverage this library for more efficient and standardized cohort definitions in their observational research.

---

**Main Finding 2: Comprehensive Metadata and Quality Assurance**

**What:** The OHDSI PL collects extensive metadata for each cohort definition, which includes user-submitted, librarian-assigned, and computer-generated metadata.

**So What:** Such metadata enhances the understandability, searchability, and reliability of the cohort definitions. Peer review processes further assure quality.

**Now What:** Researchers can have confidence in using these cohort definitions and can easily navigate the library to find definitions that suit their study requirements.

---

**Main Finding 3: Technical Infrastructure and Maintenance**

**What:** The OHDSI PL is hosted on GitHub and encapsulated within an R package, adhering to the OHDSI HADES ecosystem. It also follows a regular release cycle for updates.

**So What:** This setup ensures that the library is both accessible and maintained, with about 25 releases since the establishment of major version 3 in 2022.

**Now What:** Researchers can easily integrate the PhenotypeLibrary with other HADES packages and can expect the library to be updated and maintained regularly.
FRIDAY

Large variety
Country size
RWD data-lake

(Guy Livne, Keren Rosenstein, Atif Adam, Milou Brand, Nikolai Grewe, Ludovica Ancora, Nathan Japhet)

Large VARIETY country size DATA-LAKE

INTRO
15 years of FULL EHR data from Governmental hospitals, Birth to Death patient record.

METHODS
- Gathering Patient data from general medical center's legacy systems.
- Mapping ALL EHR data to standard OMOP concepts using machine learning pipelines and expert reviews.
- Conforming all to OMOP CDOM v5.3 structure.
- Developing ETL processes for quarterly updates.
- Applying data de-identification and privacy rules.
- Enabling OHDSI analytic tools.

RESULTS
A nationwide, 15-year RWD available for collaborative research through Kinaret's platform.

Salient features:
- Longitudinal, linked data across inpatient, outpatient, ER and specialized care.
- Diverse population covering all regions, ethnicities, religions, and socioeconomic strata in Israel.
- Predefined datasets for diseases like diabetes, heart failure, infections to accelerate research.
- Professional team to initiate studies in less than 3 months.

Contact us: https://kinaret.health.gov.il/en
Opening: Biomedical Informatics Data Scientist at Stanford

1.0 FTE  Full time  Day - 08 Hour  R2335119  Hybrid  84866 IT RESEARCH  Technology & Digital Solutions  455 Broadway, REDWOOD CITY, California

If you’re ready to be part of our legacy of hope and innovation, we encourage you to take the first step and explore our current job openings. Your best is waiting to be discovered.

Day - 08 Hour (United States of America)

This is a Stanford Health Care job.

A Brief Overview

The Biomedical Informatics Data Scientist will partner with researchers and clinicians to enable effective and efficient use of data and resources available via Stanford’s research clinical data repository (STARR) including the Electronic Health Records in the OMOP Common Data Model, radiology and cardiology imaging data and associated metadata, and new data types as they get integrated along with their databases and respective cohort query tools and interfaces e.g., OHDSI ATLAS. This individual will enable researchers to maximize their understanding, interpretation and use of these clinical and research tools for more informed and productive research, clinical trials, patient care and quality outcome projects.

Clean, extract, transform and analyze various kinds of clinical data to create analysis-ready datasets that follow the FAIR (Findable, Accessible, Interoperable and Re-usable) principles. Partner with researchers and clinicians to enable effective and efficient use of Stanford Clinical data and resources for the advancement of research and the educational mission.
The Zhang Lab at Washington University School of Medicine in St. Louis has one postdoc/senior data analyst position to work on causal machine learning and responsible AI for reliable real-world evidence generation.

more details at https://linyingzhang.com

- Postdoc:
  https://linyingzhang.com/files/Postdoc.pdf

- Data analyst:

If interested, please send CV and cover letter to linyingz@wustl.edu
Director, RWE at Gilead

Director, RWE - Data Science - OHDSI

Responsibilities:
Collaborate with researchers and data scientists to understand project requirements and translate them into OHDSI-compatible solutions. Work with databases, ensuring data integrity and optimization for OHDSI-related queries and analyses. Perform data analyses in OHDSI-related tools like ATLAS. Customize and extend OHDSI tools and applications to meet specific project needs. Collaborate with cross-functional teams to troubleshoot and resolve technical issues related to OHDSI implementations. Stay informed about OHDSI community updates, best practices, and emerging trends in observational health data research. Contribute to the development and documentation of data standards and conventions within the OHDSI community.
Where Are We Going?

Any other announcements of upcoming work, events, deadlines, etc?
Three Stages of The Journey

Where Have We Been?
Where Are We Now?
Where Are We Going?
April 16: Tools to Evaluate ETL

Frank DeFalco
Senior Director
Observational Health Data Analytics
Janssen Research & Development

Katy Sadowski
Senior Associate Director
Boehringer Ingelheim

April Olympians
Week 3 Update

Clair Blacketer
Director
Janssen Research & Development

Melanie Philofsky
Senior Business Analyst and Project Manager, Odysseus Data Services, Inc.
The weekly OHDSI community call is held every Tuesday at 11 am ET.

Everybody is invited!

Links are sent out weekly and available at:

ohdsi.org/community-calls