April Olympians #2 / Vocabulary Techniques for ETL

OHDSI Community Call
April 9, 2024 • 11 am ET
## Upcoming Community Calls

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 9</td>
<td>April Olympians Update</td>
</tr>
<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>
# Upcoming Asia-Pacific Calls

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 18</td>
<td>Community Call: Newcomers Session</td>
</tr>
<tr>
<td>May 2</td>
<td>Scientific Forum: CaRROT-Mapper Introduction and Demo</td>
</tr>
<tr>
<td>May 16</td>
<td>Community Call: Workgroup Introductions, part 1</td>
</tr>
<tr>
<td>June 6</td>
<td>Scientific Forum: OHDSI ETL/Vocabulary Mapping Tools Demo</td>
</tr>
<tr>
<td>June 20</td>
<td>Community Call: Regional Chapter Mid-Year Updates</td>
</tr>
</tbody>
</table>
Three Stages of The Journey

Where Have We Been?
Where Are We Now?
Where Are We Going?
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>Generative AI and Analytics</td>
</tr>
<tr>
<td>Tuesday</td>
<td>3 pm</td>
<td>OMOP CDM Oncology WG- Outreach/Research Subgroup</td>
</tr>
<tr>
<td>Wednesday</td>
<td>9 am</td>
<td>Patient-Level Prediction</td>
</tr>
<tr>
<td>Wednesday</td>
<td>2 pm</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>Wednesday</td>
<td>3 pm</td>
<td>Joint Vulcan/OHDS Meeting</td>
</tr>
<tr>
<td>Thursday</td>
<td>9:30 am</td>
<td>Network Data Quality</td>
</tr>
<tr>
<td>Thursday</td>
<td>12 pm</td>
<td>Strategus HADES Subgroup</td>
</tr>
<tr>
<td>Thursday</td>
<td>7 pm</td>
<td>Dentistry</td>
</tr>
<tr>
<td>Friday</td>
<td>9 am</td>
<td>Phenotype Development and Evaluation</td>
</tr>
<tr>
<td>Friday</td>
<td>10 am</td>
<td>GIS-Geographic Information System</td>
</tr>
<tr>
<td>Friday</td>
<td>11:30 am</td>
<td>Clinical Trials</td>
</tr>
<tr>
<td>Friday</td>
<td>11:30 am</td>
<td>Steering Group</td>
</tr>
<tr>
<td>Friday</td>
<td>10 pm</td>
<td>China Chapter</td>
</tr>
<tr>
<td>Monday</td>
<td>10 pm</td>
<td>Africa Chapter</td>
</tr>
<tr>
<td>Monday</td>
<td>11 am</td>
<td>Data Bricks User Group</td>
</tr>
<tr>
<td>Monday</td>
<td>2 pm</td>
<td>Electronic Animal Health Records</td>
</tr>
</tbody>
</table>
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 CBER BEST Initiative Seminar Series is designed to share and discuss recent research of relevance to ongoing and future surveillance activities of CBERT regulated products, namely biologics. The series focuses on safety and effectiveness of biologics including vaccines, blood components, blood-derived products, tissues and advanced therapies. The seminars will provide information on characteristics of biologics, required infrastructure, study designs, and analytic methods utilized for pharmacovigilance and pharmacoepidemiologic studies of biologics. They will also cover information regarding potential data sources, informatics challenges and requirements, 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 from outside FDA.

Below you will find details of upcoming CBERT BEST 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.

@OHDSI
www.ohdsi.org
#JoinTheJourney

Upcoming Seminars

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

Previous Seminars

- Jan. 17, 2024 - Anna Ostropoleva, Oxysense Data Services
- Dec. 6, 2023 - Jenny Sun, Pfizer
- June 14, 2023 - Katsurayna Bykova, Harvard Medical School
- May 1, 2023 - Xintong Li and Daniel Prieto-Alhambra, University of Oxford, NIDRMS
- Apr. 12, 2023 - Kielej Boliart, P-RX
- Mar. 23, 2023 - Marilt Schum, Janssen R&D
- Feb. 8, 2023 - Fan Du, UCLA

ohdsi.org/cber-best-seminar-series
<table>
<thead>
<tr>
<th>Package</th>
<th>Version</th>
<th>Maintainer(s)</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achilles</td>
<td>v1.7.2</td>
<td>Frank DeFalco</td>
<td>CRAN</td>
</tr>
<tr>
<td>Andromeda</td>
<td>v0.6.0</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
</tr>
<tr>
<td>BigData</td>
<td>v1.0.2</td>
<td>Martijn Schuemie</td>
<td>GitHub</td>
</tr>
<tr>
<td>ResNetAdaptiveRidge</td>
<td>v1.0.0</td>
<td>Marc Sachard</td>
<td>CRAN</td>
</tr>
<tr>
<td>Cotor</td>
<td>v2.0.0</td>
<td>Martin Lavallée</td>
<td>GitHub</td>
</tr>
<tr>
<td>Characterization</td>
<td>v0.1.0</td>
<td>Jenna Reps</td>
<td>GitHub</td>
</tr>
<tr>
<td>Circel</td>
<td>v1.3.3</td>
<td>Chris Knoll</td>
<td>GitHub</td>
</tr>
<tr>
<td>CohortDiagnostics</td>
<td>v2.2.0</td>
<td>Jamie Gilbert</td>
<td>GitHub</td>
</tr>
<tr>
<td>CohortExplorer</td>
<td>v0.1.0</td>
<td>Gowtham Rao</td>
<td>CRAN</td>
</tr>
<tr>
<td>CohortGenerator</td>
<td>v0.3.1</td>
<td>Anthony Sena</td>
<td>GitHub</td>
</tr>
<tr>
<td>CohortMethod</td>
<td>v5.2.1</td>
<td>Martijn Schuemie</td>
<td>GitHub</td>
</tr>
<tr>
<td>Cyclopes</td>
<td>v3.4.0</td>
<td>Marc Sachard</td>
<td>CRAN</td>
</tr>
<tr>
<td>DatabaseConnector</td>
<td>v0.3.1</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
</tr>
<tr>
<td>DataQualityDashboard</td>
<td>v2.6.0</td>
<td>Katy Sadowski</td>
<td>GitHub</td>
</tr>
<tr>
<td>DeepPatientLevelPrediction</td>
<td>v2.0.3</td>
<td>Egill Fridgeirsson</td>
<td>GitHub</td>
</tr>
<tr>
<td>EmpiricalCalibration</td>
<td>v3.1.2</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
</tr>
<tr>
<td>EnsemblePatientLevelPrediction</td>
<td>v1.0.2</td>
<td>Jenna Reps</td>
<td>GitHub</td>
</tr>
<tr>
<td>Evocemia</td>
<td>v1.0.3</td>
<td>Frank DeFalco</td>
<td>GitHub</td>
</tr>
<tr>
<td>EvidenceSynthesis</td>
<td>v0.5.0</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
</tr>
<tr>
<td>FeatureExtraction</td>
<td>v2.4.1</td>
<td>Anthony Sena</td>
<td>GitHub</td>
</tr>
<tr>
<td>Hydra</td>
<td>v0.4.0</td>
<td>Anthony Sena</td>
<td>GitHub</td>
</tr>
<tr>
<td>IterativeHardThresholding</td>
<td>v1.0.2</td>
<td>Marc Suchard</td>
<td>CRAN</td>
</tr>
<tr>
<td>MethodEvaluation</td>
<td>v2.3.0</td>
<td>Martijn Schuemie</td>
<td>GitHub</td>
</tr>
<tr>
<td>OhdsiSharing</td>
<td>v0.2.2</td>
<td>Lee Evans</td>
<td>GitHub</td>
</tr>
<tr>
<td>OhdsiShinyModules</td>
<td>v2.1.2</td>
<td>Jenna Reps</td>
<td>GitHub</td>
</tr>
<tr>
<td>ParallelLogger</td>
<td>v3.3.0</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
</tr>
<tr>
<td>PatientLevelPrediction</td>
<td>v6.3.7</td>
<td>Jenna Reps &amp; Peter Rijnbeek</td>
<td>GitHub</td>
</tr>
<tr>
<td>PhenotypeLibrary</td>
<td>v3.32.0</td>
<td>Gowtham Rao</td>
<td>GitHub</td>
</tr>
<tr>
<td>PheEvaluator</td>
<td>v2.2.31</td>
<td>Joel Swerdel</td>
<td>GitHub</td>
</tr>
<tr>
<td>ResultModelManager</td>
<td>v0.5.6</td>
<td>Jamie Gilbert</td>
<td>GitHub</td>
</tr>
<tr>
<td>ROhdsiWebApi</td>
<td>v1.3.3</td>
<td>Gowtham Rao</td>
<td>GitHub</td>
</tr>
<tr>
<td>SelfControlledCaseSeries</td>
<td>v0.1.1</td>
<td>Martijn Schuemie</td>
<td>GitHub</td>
</tr>
<tr>
<td>SelfControlledCohort</td>
<td>v1.5.0</td>
<td>Jamie Gilbert</td>
<td>GitHub</td>
</tr>
<tr>
<td>ShinyAppBuilder</td>
<td>v2.0.1</td>
<td>Jenna Reps</td>
<td>GitHub</td>
</tr>
<tr>
<td>SqlRender</td>
<td>v1.17.0</td>
<td>Martijn Schuemie</td>
<td>CRAN</td>
</tr>
</tbody>
</table>

@OHDSI
www.ohdsi.org

#JoinTheJourney
Spotlight: Melanie Philofsky

Melanie Philofsky is a Senior Business & Data Analyst with Odysseus Data Services, Inc. She is responsible for the harmonization of various healthcare data sources into the OMOP Common Data Model to support research endeavors. Her areas of expertise include clinical informatics, data analysis, data quality, ETL conversions, EHR data, the OMOP CDM and data modeling of new domains.

Prior to earning her MS in Healthcare Informatics, she was an ICU RN. She knows and understands the clinical workflow and UI of an EHR system to the backend where data is pulled for transformation to the OMOP CDM. She was the 2022 Titan Award honoree for Contributions in Data Standards.

In the latest edition of the Collaborator Spotlight, Melanie discusses her career journey, her work with the Healthcare Systems and Themis workgroups, plans for the April Olympians Collab-a-thon, and more!

Can you discuss your career journey and why you transitioned from nursing to your work in health data?

As a bedside RN in the ICU, I frequently researched journal articles and practice guidelines to find evidence in support of nursing practice and theory. I hungered for more and better information and knowledge to provide the best, scientifically supported practices to holistically care for my patients and their families. It was always my intention to continue my education and earn an advanced degree. When I started researching career pathways for nurses, I came upon informatics. The more I learned about this field, the more I saw myself at the intersection of science and data to extract actionable information, knowledge, and wisdom to positively influence patient care. In the ICU I would care for 1 or 2 people at a time. With observational research, I am supporting hundreds to millions of people in their health journey by producing evidence for them to make informed decisions.

ohdsi.org/spotlight-melanie-philofsky
DevCon 2024: April 26, 9 am-3 pm ET

The third annual OHDSI DevCon will be held virtually on Friday, April 26, from 9 am-3 pm ET.

Join leaders from our Open-Source Community for a day to both welcome and inform both new and veteran developers within the OHDSI Community.
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
Implementing the OMOP common data model in an NHS Trust using DBT

(Quinta Ashcroft, Timothy Howcroft, Dale Kirkwood, Jo Knight, Vishnu V Chandrabalan)

Implementing the OMOP common data model using dbt
Quin Ashcroft, Dale Kirkwood, Tim Howcroft, Jo Knight, Stephen Dobson, Vishnu V Chandrabalan

MONDAY

Background
Lancashire Teaching Hospitals NHS Foundation Trust (LTH) is a digitally mature secondary care provider, major trauma centre and multi-speciality hospital, covering a large catchment in Northwest England and part of the UK National Health Service. LTH have routinely collected healthcare data for more than 1.7 million patients spanning over 15 years, covering most aspects of secondary care.

Electronic health data collected using a primary EPR and multiple disparate specialist clinical systems are held in isolated, poorly documented databases from multiple vendors with no straightforward method to create a single, linked, person-centric, semantic view. The OMOP Common Data Model was chosen for its person-centric design, rich OMOP analytic software ecosystem, vibrant global research community and opportunities for national and international collaboration to accelerate research as well as to support near real-time operational and clinical intelligence to drive transformation and improvements to patient care pathways and organisational efficiency.

We describe the use of dbt (data build tool) to implement a complex extract-load-transform (ELT) workflow that transforms data from multiple sources daily and incrementally, into a single OMOP database.

Methods
ELT (extract-load-transform) workflow with dbt engine was adopted to implement a complete ELT development for several reasons:
- Collaborative development as a team with version control
- Improved data visibility, maintenance and automation
- Multiple target architectures and parallelism
- Auto-documenting with pipelines as directed acyclic graphs (DAG) enriched by information for development, testing, and audit.
- Ability to build sections of the DAG by selecting modules or tasks
- Integration with Prefet for workflow orchestration

Incremental loading for selected modules using custom strategies minimised computational load on source systems, reduced build times and made daily updates feasible.

Vocabulary mapping of source concepts to standardised vocabularies (SNOMED, ICD10, OPUS) from Athena was done using json. These were exported to multiple domain-specific flat files which were re-ordered to version control and incorporated into dbt as seeds. A single module json file per concept file was created as a union of these seed modules and integrated into downstream pipeline.

DBT enabled rapid, collaborative transformation of data from multiple, disparate source systems in Oracle, Sybase and SQL Server into OMOP in SQL Server.

Data from multiple sources were harmonised in a staging layer (Fig. 1) before being separated by domain into OMOP tables. Figure 3 shows the timeline for each month emphasising the importance of incremental updates.

Conclusions
The use of dbt and git allowed for additional sources to be integrated into existing pipelines with minimal effort and enabled a clear path for integration of future data sources.

The documentation and DAG data lineage generated by dbt were published online and shared with regional development partners.

References
1. dbt [https://dbt.readthedocs.io]
2. Prefet [https://prefet.readthedocs.io]
3. dbt docs: [https://dbt.readthedocs.io]
4. SQL for OMOP: [https://dbt.readthedocs.io]
Mining Data Outside the Box: Internet as a New Source for Common Data Model

(Min-Gyu Kim, Min ho An, GyuBeom Hwang, Rae Woong Park)

**Background**

While the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) standardizes data acquired in healthcare settings, EMR data is not the only source of healthcare data. The Internet, such as social media, patient forums, and other online sources can also be a valuable source of real-world health data.

However, Internet data is not as easy to handle as CDM. It is often unstructured and can be difficult to extract meaningful information from.

In this paper, we present our first step in extracting and formatting medical data mined from the Internet into OMOP-CDM. A certain degree of deduction is necessary to use texts from the Internet as a source to feed OMOP-CDM. To tackle this problem, we used a generative large language model (LLM) to generate text about the logical flow of extraction.

**Methods**

We focused on extracting the date of diagnosis from posts submitted by diabetes patients on the Internet community “Reddit”. We designed a method consisting of two steps: first, we used text generation models to create text explaining why the date of diagnosis is estimated as such; second, we evaluated the output in three aspects: factual, logical, mathematical and formatting correctness.

We used LLAMA-30b supercot, a variation of the large language model (LLM) “Large Language Model Meta AI” (LLAMA) from Meta. The LLM extracts information related to the date of diagnosis for diabetes mentioned in the posts, and answers with the estimated date of diagnosis. It was explicitly asked to include the reasoning about how it produced that date.

**Results**

While the outputs generated, only 4 of them included factual inaccuracies. Furthermore, when focusing specifically on the 23 post submissions that provided context regarding the date of diagnosis, none of the outputs were factually incorrect. However, in terms of logical deductions, out of the same 23 post submissions, 18 outputs were logically correct while 5 were deemed incorrect.

**Conclusions**

This paper suggests the potential of generative language models being utilized in mining medical data from the Internet and formatting them for convenient usage. At the moment, its accuracy is not optimal yet. Nonetheless, our work shows the feasibility of building CDM out of a data source that is not a part of the healthcare system. We believe similar approaches could be used on a variety of Internet data sources and conventional EMR alike. With the development of additional modules to assist LLM, the Internet may become a new source of medical data to feed OMOP-CDM.

**Acknowledgment**

- This research was funded by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR18C0001).
- This research was supported by a Government-wide R&D Fund project for infectious disease research (G4F), Republic of Korea (grant number: H20200034).

Contact: mairjrnms@gmail.com
Forecasting Daily Incidence of Respiratory Symptoms: A Comparative Study on Time Series Models using OMOP-CDM in South Korea

(min Ho An, Min-Gyu Kim, GyuBeom Hwang, ByungJin Choi, Rae Woong Park)

**Background**
- With the outbreak of the COVID-19 pandemic, the significance of infectious disease surveillance and quarantine prediction has been emphasized. Several reports associated with prediction of respiratory infectious disease including COVID-19 have been published.
- Respiratory infectious disease like COVID-19 can disseminate rapidly, given the impossibility of restricting respiratory activities. To monitor disease spread, four district hospitals in South Korea recently began collaborating to collect data using the Observational Medical Outcomes Partnership - Common Data Model (OMOP-CDM) under the project named PHAKOS (Platform for Harmonizing and Accessing Data in Real-time Infectious Disease Surveillance).
- During its nascent developmental stage in this project, we sought to compare two potent models: ARIMA and Prophet, to predict the daily occurrence of respiratory symptoms. This study aims to assess each model’s effectiveness and verify their accuracy in predicting the daily incidence of respiratory symptoms.

**Methods**
- Patients visited or admitted to the emergency or infectious disease department presenting with symptoms including fever, cough, or cough at Ajou University Hospital in South Korea were defined as respiratory symptom-related visits.
- A total of 18,839 visits with respiratory symptoms were recorded from January 1, 2018, to December 31, 2021.
- The primary outcome in this study was the daily occurrence of respiratory symptoms classified above. To forecast this, we employed two models: ARIMA and Prophet.
- The total dataset was divided into train and test data, first allotting 80% towards the training set to build the model. The remaining 20% of the data was reserved as a test set to evaluate the model's predictive accuracy. All analyses were performed via Python 3.7.

**Conclusions**
- In the task of predicting daily counts of respiratory symptoms in South Korea, the ARIMA and Prophet models mostly preserved forecasts within a 95% confidence interval.
- Despite ARIMA's superior accuracy, denoted by a lower MAE and RMSE, the Prophet model offered a more realistic reflection of the day's variance.
- Therefore, model selection hinges on the study's specific objectives: ARIMA for numerical precision, and Prophet for discerning variance and trend changes.
- This study emphasizes the imperative of additional research to refine these models, enhancing infectious disease surveillance—a key component of healthcare preparedness in pandemic scenarios.

**Acknowledgement**
- This research was funded through a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C004).
Title: Observational Research in Dentistry
A Scoping Review

**PRESENTER:** Robert Koski

**INTRO**
The aims for the scoping review are to:
1. Describe observational research implementations and challenges in dentistry, and
2. Describe characteristics of successful implementations of observational research in healthcare

**METHODS**
1. Following the PRISMA-SuR protocol for scoping reviews
2. Interviewing subject matter experts
3. Conducting searches in PubMed and Scopus
4. Screening articles based on inclusion/exclusion criteria

**Inclusion criteria:**
- Use patient-level data from multiple sources
- Use or discuss a common data model or standardized terminology
- Discuss the implications, challenges, and attempts to conduct observational research in dentistry
- Discuss implementation of a common data model in a given healthcare setting or specialty

**Exclusion criteria:**
- Article published before 2010
- Article is not related to observational research
- Article does not pertain to the process of conducting observational research with health data
- Letters to the editor, editorials, critical reviews

Observational research can help explore the link between oral health and systemic disease.

The Dentistry Workgroup is addressing barriers to observational research on oral health.

(Robert Koski, Danielle Boyce, Brock Johnson, Adam Bouras, Swetha Kiranmayi Jakkuva)
Integration of Scalable Natural Language Processing to the Atlas Cohort Building Workflow

(Pavan Parimi, Selvin Soby, Pavel Gorjacko, Chandra Nelapati, Boudewijn Aasman, Manuel Wahlle, Reetam Nath, Parsa Mirhaji)

**BACKGROUND**

Integration of scalable natural language processing (NLP) and cohort-based analysis on observational, electronic health record (EHR) data to identify patients from clinical data with diabetes (eDiabetes) for research. This process provides a mechanism for researchers to build alternate cohorts from data that are not included in an atlas that was developed to build differential cohorts. The clinical text using NLP engines (EINSTEIN and Elastic Search). Clinical text analysis and knowledge extraction from EHR data was performed using the CDEMS platform, an open-source natural language processing system. The platform was linked to the clinical cohort building workflow and the components were specifically designed for the clinical data, based on user needs and recent advances in specific entities, relationships between those entities, part-of-speech tagging, and dependency parsing.

**METHODS**

A: Data preprocessing: The source database containing the clinical text is obtained. The data is then preprocessed by understanding the various sections within the clinical structure, such as impression, plan, and data by data information.

B: Dictionary integration and Elasticsearch storage: The CDEMS software is configured, including the integration of the UMLS and CDEMS dictionaries, relation extraction, negation and context extraction, and any other required components. The output from CDEMS and other analysis engines, along with the related data, is serialized in JSON-LD (JSON for Linked Data) and integrated into an Elastic Search cluster for storage.

C: Additional Analytical Processing: To perform tasks like named entity recognition, part-of-speech tagging, and information extraction, a pipeline for knowledge extraction that uses a large language model is configured to produce information in clinical text. The final text, metadata, and annotations generated by the analytic pipelines are then visualized in an Elastic Search database. Finally, necessary indexing is implemented to facilitate efficient retrieval of the preprocessed data.

D: User Interaction Model: A specific user experience and interaction model was developed to expose cohorts generated or shared via Atlas to the NLP engine for just-in-time querying. A variable cohort-based query engine was developed to enable submitting concise pattern search queries or templates based on user-defined phrases. The system allows users to customize cohort terms and enrich the cohort with user-defined metadata, contributing being able to distinguish between terms of people and conceptual terminology. A dashboard has been developed to visualize cohort terms using interactive visualizations from the atlas view.

**RESULTS**

As of October 2023, over 130,000 notes have been tagged and annotated by the NLP pipeline for the OHDSI Atlas, and over 2,000 cohort terms have been created. The system allows for the customization of cohort terms and enriches the cohort with user-defined metadata. The dashboard provides a way to distinguish between terms of people and conceptual terminology. A dashboard has been developed to visualize cohort terms using interactive visualizations from the atlas view.

**CONCLUSION**

The NP engine that we developed and implemented further has proven to use the ability to handle large electronic health record data from the OHDSI platform as well as semantic annotation and matching capabilities. The system provides a way to enrich and manage cohort terms, making it easier to identify and extract relevant information from EHR data. The dashboard provides an interactive way to view cohort terms and enrich them with user-defined metadata.

**REFERENCES**


**Figure 1:** Architecture of the NLP pipeline and cohort-based analysis on EHR data.

**Figure 2:** Text-based query builder for cohort building.

**Figure 3:** Example clinical text with annotations, mappings, and deidentification.
Opening: Biomedical Informatics Data Scientist at Stanford

Biomedical Informatics Data Scientist
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
  o Postdoc: https://linyingzhang.com/files/Postdoc.pdf
  o Data analyst: https://linyingzhang.com/files/Analyst.pdf

• 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?
OHDSI Workgroup
Objectives and Key Results (OKR)

Rehabilitation Workgroup
WG Name: Rehabilitation Workgroup
WG Leads: Esther Janssen & Ruud Selles

Mission statement

Promote better rehabilitation care by leveraging the OHDSI collaborative to enable large scale observational rehabilitation research
WG Name: Rehabilitation Workgroup  
WG Leads: Esther Janssen & Ruud Selles

1. Objective 1: Create awareness of OHDSI in rehabilitation research and build a learning community

2024 Key goals/results:

1. Establish a minimum of 6 workgroup meetings
2. Have at least 50 working group members
3. Increase international awareness of what OHDSI and OMOP-CMD can provide in the rehabilitation research community through social media, presentations, and meetings
WG Name: Rehabilitation Workgroup
WG Leads: Esther Janssen & Ruud Selles

1. Objective 2: Identify challenges and find best practices in using OMOP-CDM for rehabilitation research data

2024 Key goals/results:

1. Identify and define challenges in mapping rehabilitation-specific outcome data to the OMOP-CDM (e.g., PROMS)
2. Identify and define challenges in mapping rehabilitation-specific treatments to the OMOP-CDM (e.g., complex treatments, multidisciplinary treatments)
3. Develop best practices in mapping rehabilitation-specific data to the OMOP-CDM
4. Reach out to other working groups (e.g., CMD, psychiatry) and OHDSI members to discuss our challenges and possible solutions
WG Name: Rehabilitation Workgroup
WG Leads: Esther Janssen & Ruud Selles

1. Objective 3: Initiate a StudyAthon as a proof of concept for the value of OHDSI in rehabilitation science

2024 Key goals/results:

1. Identify a list of topics for a network study with two or more international partners as a proof of concept and a community learning experience
2. Perform the StudyAthon at the end of 2024 or in 2025
April 9: Vocabulary Techniques for ETL

**Alexander Davydov**
Director, Lead of Medical Ontologies
Odysseus Data Services, Inc.

**Dmitry Dymshyts**
Associate Director
Janssen Research & Development

**Tanya Skugarevskaya**
Vocabulary Team
Odysseus Data Services, Inc.

**Anna Ostropolets**
Associate Director
Janssen Research & Development

---

**April Olympians**
**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: ohdssi.org/community-calls