April Olympians #2 / Vocabulary Techniques for ETL

OHDSI Community Call
April 9, 2024 • 11 am ET

Upcoming Community Calls

Date	Topic
April 9	April Olympians Update | Presentation: Vocabulary for ETL
April 16	April Olympians Update | Presentation: Tools to Evaluate ETL
April 23	April Olympians Update | Presentation: Themis & CDM Process Overview
April 30	April Olympians Update | Presentation: What We Achieved & How You Can Use It
May 7	DevCon 2024 Review
May 14	10-Minute Tutorials
May 21	Open Studies in the OHDSI Community
May 28	Collaborator Showcase Brainstorm
June 4	NO CALL – EUROPEAN SYMPOSIUM
June 11	European Symposium Review
June 18	Application of LLMs In Evidence Generation Process
June 25	Recent OHDSI Publications

Upcoming Asia-Pacific Calls

Date	Topic
April 18	Community Call: Newcomers Session
May 2	Scientific Forum: CaRROT-Mapper Introduction and Demo
May 16	Community Call: Workgroup Introductions, part 1
June 6	Scientific Forum: OHDSI ETL/Vocabulary Mapping Tools Demo
June 20	Community Call: Regional Chapter Mid-Year Updates

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Pawel Rajwa, Angelika Borkowetz, Thomas Abbott, Andrea Alberti, Anders Bjartell, James T. Brash, Riccardo Campi, Andrew Chilelli, Mitchell Conover, Niculae Constantinovici, Eleanor Davies, Bertrand De Meulder, Sherrine Eid, Mauro Gacci, Asieh Golozar, Haroon Hafeez, Samiul Haque, Ayman Hijazy, Tim Hulsen, Andreas Josefsson, Sara Khalid, Raivo Kolde, Daniel Kotik, Samu Kurki, Mark Lambrecht, Chi-Ho Leung, Julia Moreno, Rossella Nicoletti, Daan Nieboer, Marek Oja, Soundarya Palanisamy, Peter Prinsen, Christian Reich, Giulio Raffaele Resta, Maria J Ribal, Juan Gómez Rivas, Emma Smith, Robert Snijder, Carl Steinbeisser, Frederik Vandenberghe, Philip Cornford, Susan Evans-Axelsson, James N'Dow, and Peter-Paul M Willemse** on the publication of **Research Protocol for an Observational Health Data Analysis on the Adverse Events of Systemic Treatment in Patients with Metastatic Hormone-sensitive Prostate Cancer: Big Data Analytics Using the PIONEER Platform** in *European Urology Open Science.*


EUROPEAN UROLOGY OPEN SCIENCE 63 (2024) 81–88

available at www.sciencedirect.com
journal homepage: www.eu-openscience.europeanurology.com

European Association of Urology

Trial Protocol

Research Protocol for an Observational Health Data Analysis on the Adverse Events of Systemic Treatment in Patients with Metastatic Hormone-sensitive Prostate Cancer: Big Data Analytics Using the PIONEER Platform

Pawel Rajwa[a,b], Angelika Borkowetz[c], Thomas Abbott[d], Andrea Alberti[e], Anders Bjartell[f], James T. Brash[g], Riccardo Campi[e], Andrew Chilelli[h], Mitchell Conover[i], Niculae Constantinovici[j], Eleanor Davies[g], Bertrand De Meulder[k], Sherrine Eid[l], Mauro Gacci[e], Asieh Golozar[m,n], Haroon Hafeez[o], Samiul Haque[l], Ayman Hijazy[k], Tim Hulsen[p], Andreas Josefsson[q,r], Sara Khalid[s], Raivo Kolde[t], Daniel Kotik[u,v], Samu Kurki[w], Mark Lambrecht[l], Chi-Ho Leung[x], Julia Moreno[l], Rossella Nicoletti[e], Daan Nieboer[y], Marek Oja[t], Soundarya Palanisamy[l], Peter Prinsen[z], Christian Reich[m,n], Giulio Raffaele Resta[e], Maria J. Ribal[aa], Juan Gómez Rivas[bb], Emma Smith[cc], Robert Snijder[i], Carl Steinbeisser[dd], Frederik Vandenberghe[l], Philip Cornford[ee], Susan Evans-Axelsson[j], James N'Dow[ff], Peter-Paul M. Willemse[gg,]*

[a] Department of Urology, Medical University of Silesia, Zabrze, Poland; [b] Department of Urology, Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria; [c] Department of Urology, University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany; [d] European Association of Urology, Nijmegen, The Netherlands; [e] Unit of Urological Robotic Surgery and Renal Transplantation, University of Florence, Careggi Hospital, Florence, Italy; [f] Department of Translational Medicine, Lund University, Lund, Sweden; [g] IQVIA, Real World Solutions, Brighton, UK; [h] Astellas Pharma Europe Ltd, Surrey, UK; [i] Janssen Research & Development, Titusville, NJ, USA; [j] Bayer AG, Berlin, Germany; [k] Association EISBM, Vourles, France; [l] SAS Institute, Cary, NC, USA; [m] Odysseus Data Services, New York, NY, USA; [n] OHDSI Center, Northeastern University, Boston, MA, USA; [o] Shaukat Khanum Memorial Cancer Hospital & Research Centre, Peshawar, Pakistan; [p] Department of Hospital Services & Informatics, Philips Research, Eindhoven, The Netherlands; [q] Department of Urology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; [r] Wallenberg Center for Molecular Medicine, Umeå University, Umeå, Sweden; [s] University of Oxford, Oxford, UK; [t] Institute of Computer Science, University of Tartu, Tartu, Estonia; [u] Center for Advanced Systems Understanding, Görlitz, Germany; [v] Helmholtz-Zentrum Dresden-Rossendorf, Dresden, Germany; [w] Bayer OY, Turku, Finland; [x] S.H. Ho Urology Centre, Department of Surgery, The Chinese University of Hong Kong, Hong Kong, China; [y] Erasmus MC University Medical Center, Rotterdam, The Netherlands; [z] Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands; [aa] Uro-Oncology Unit, Hospital Clinic, University of Barcelona, Barcelona, Spain; [bb] Department of Urology, Hospital Clinico San Carlos, Madrid, Spain; [cc] Guidelines Office, European Association of Urology, Arnhem, The Netherlands; [dd] Collaborate Project Management, Munich, Germany; [ee] Liverpool University Hospitals NHS Trust, Liverpool, UK; [ff] Academic Urology Unit, University of Aberdeen, Aberdeen, UK; [gg] Department of Urology, Cancer Center, University Medical Center Utrecht, Utrecht, The Netherlands

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Tuesday	12 pm	Generative AI and Analytics
Tuesday	3 pm	OMOP CDM Oncology WG- Outreach/Research Subgroup
Wednesday	9 am	Patient-Level Prediction
Wednesday	2 pm	Natural Language Processing
Wednesday	3 pm	Joint Vulcan/OHDS Meeting
Thursday	9:30 am	Network Data Quality
Thursday	12 pm	Strategus HADES Subgroup
Thursday	7 pm	Dentistry
Friday	9 am	Phenotype Development and Evaluation
Friday	10 am	GIS-Geographic Information System
Friday	11:30 am	Clinical Trials
Friday	11:30 am	Steering Group
Friday	10 pm	China Chapter
Monday	10 am	Africa Chapter
Monday	11 am	Data Bricks User Group
Monday	2 pm	Electronic Animal Health Records

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.

ohdsi.org/cber-best-seminar-series

Next CBER BEST Seminar: Apr. 17

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Who We Are | Updates & News | Standards | Software Tools | Network Studies | Community Forums | Education | New To OHDSI?
Community Calls | Past Events | Workgroups | 2023 'Our Journey' Annual Report | This Week In OHDSI | Support & Sponsorship
CBER Best Seminars | 2024 Europe Symposium | 2024 Global Symposium | Github | YouTube | Twitter | LinkedIn | Newsletters

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 CBER 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.

BEST

Below you will find details of upcoming CBER 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.

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 CBER 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.

BEST

Below you will find details of upcoming CBER 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.

Upcoming Seminars

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

Previous Seminars

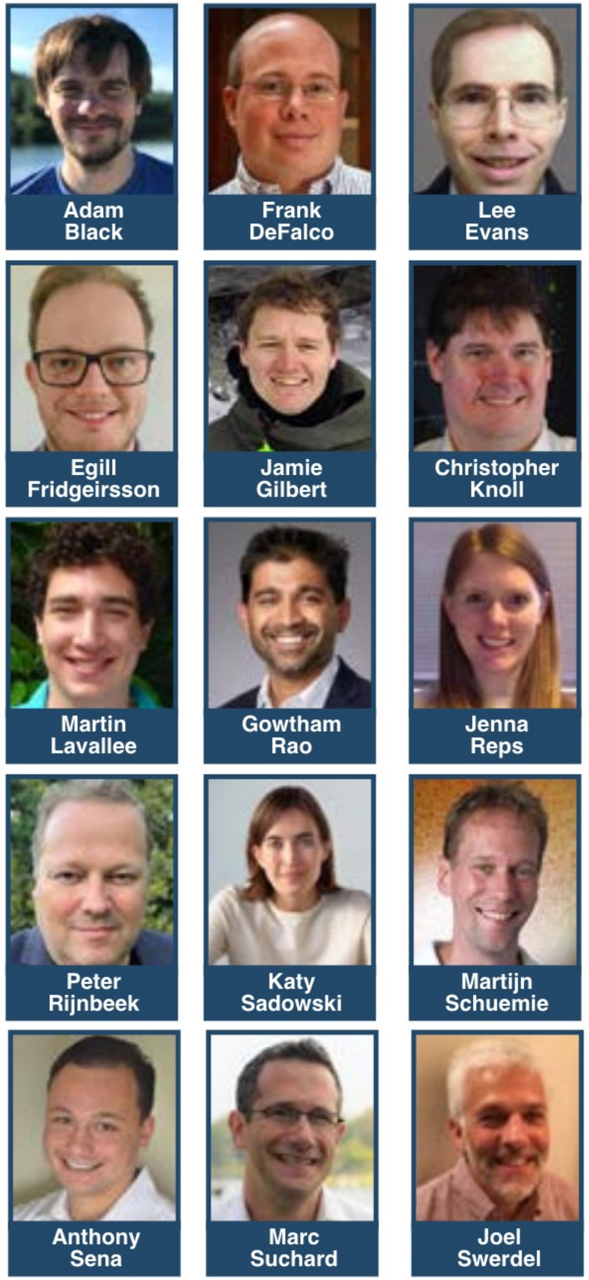
+ Jan. 17, 2024 · Anna Ostropolets, Odysseus Data Services
+ Dec. 6, 2023 · Jenny Sun, Pfizer
+ June 14, 2023 · Katsiaryna Bykov, Harvard Medical School
+ May 3, 2023 · Xintong Li and Daniel Prieto-Alhambra, University of Oxford, NDORMS
+ Apr. 12, 2023 · Kaatje Bollaerts, P-95
+ Mar. 22, 2023 · Martijn Schuemie, Janssen R&D
+ Feb. 8, 2023 · Fan Bu, UCLA

ohdsi.org/cber-best-seminar-series

HADES-wide Release 2024Q1

HADES | Packages | Validation | Publications | Support

Package	Version	Maintainer(s)	Availability
Achilles	v1.7.2	Frank DeFalco	CRAN
Andromeda	v0.6.6	Martijn Schuemie	CRAN
BigKnn	v1.0.2	Martijn Schuemie	GitHub
BrokenAdaptiveRidge	v1.0.0	Marc Suchard	CRAN
Capr	v2.0.7	Martin Lavallee	GitHub
Characterization	v0.1.5	Jenna Reps	GitHub
CirceR	v1.3.2	Chris Knoll	GitHub
CohortDiagnostics	v3.2.5	Jamie Gilbert	GitHub
CohortExplorer	v0.1.0	Gowtham Rao	CRAN
CohortGenerator	v0.8.1	Anthony Sena	GitHub
CohortMethod	v5.2.1	Martijn Schuemie	GitHub
Cyclops	v3.4.0	Marc Suchard	CRAN
DatabaseConnector	v6.3.2	Martijn Schuemie	CRAN
DataQualityDashboard	v2.6.0	Katy Sadowksi	GitHub
DeepPatientLevelPrediction	v2.0.3	Egill Fridgeirsson	GitHub
EmpiricalCalibration	v3.1.2	Martijn Schuemie	CRAN
EnsemblePatientLevelPrediction	v1.0.2	Jenna Reps	GitHub
Eunomia	v1.0.3	Frank DeFalco	GitHub
EvidenceSynthesis	v0.5.0	Martijn Schuemie	CRAN
FeatureExtraction	v3.4.1	Anthony Sena	GitHub
Hydra	v0.4.0	Anthony Sena	GitHub
IterativeHardThresholding	v1.0.2	Marc Suchard	CRAN
MethodEvaluation	v2.3.0	Martijn Schuemie	GitHub
OhdsiSharing	v0.2.2	Lee Evans	GitHub
OhdsiShinyModules	v2.1.2	Jenna Reps	GitHub
ParallelLogger	v3.3.0	Martijn Schuemie	CRAN
PatientLevelPrediction	v6.3.7	Jenna Reps & Peter Rijnbeek	GitHub
PhenotypeLibrary	v3.32.0	Gowtham Rao	GitHub
PheValuator	v2.2.11	Joel Swerdel	GitHub
ResultModelManager	v0.5.6	Jamie Gilbert	GitHub
ROhdsiWebApi	v1.3.3	Gowtham Rao	GitHub
SelfControlledCaseSeries	v5.1.1	Martijn Schuemie	GitHub
SelfControlledCohort	v1.6.0	Jamie Gilbert	GitHub
ShinyAppBuilder	v2.0.1	Jenna Reps	GitHub
SqlRender	v1.17.0	Martijn Schuemie	CRAN



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?

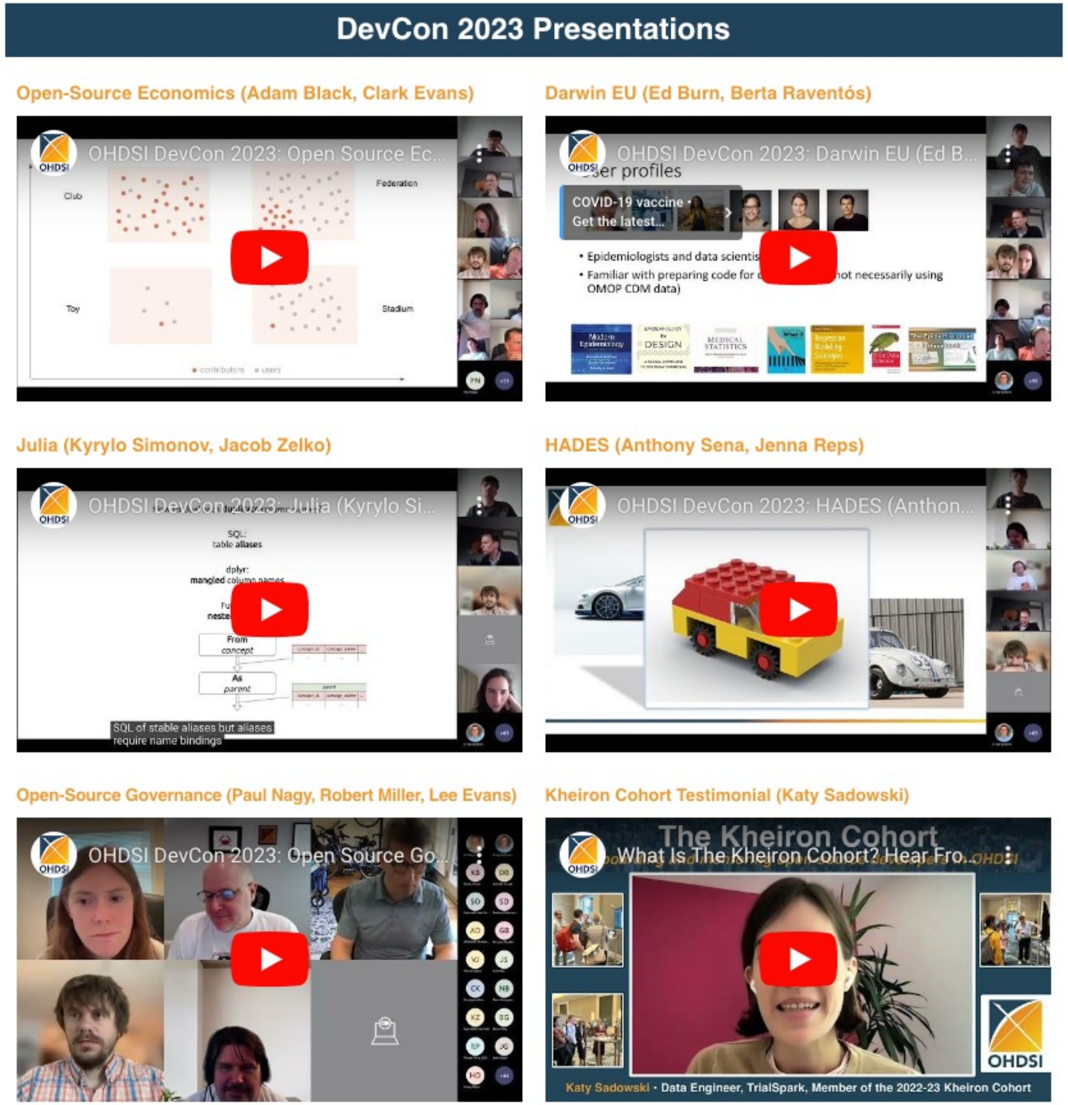
As a bedside RN in the ICU, I frequently researched journal articles and practice guidelines to find evidence in support of nursing practice and theory. I hungered for more and better information and knowledge to provide the best, scientifically supported practices to holistically care for my patients 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 hundred 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.



OHDSI Global Symposium

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



OHDSI Europe Symposium

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

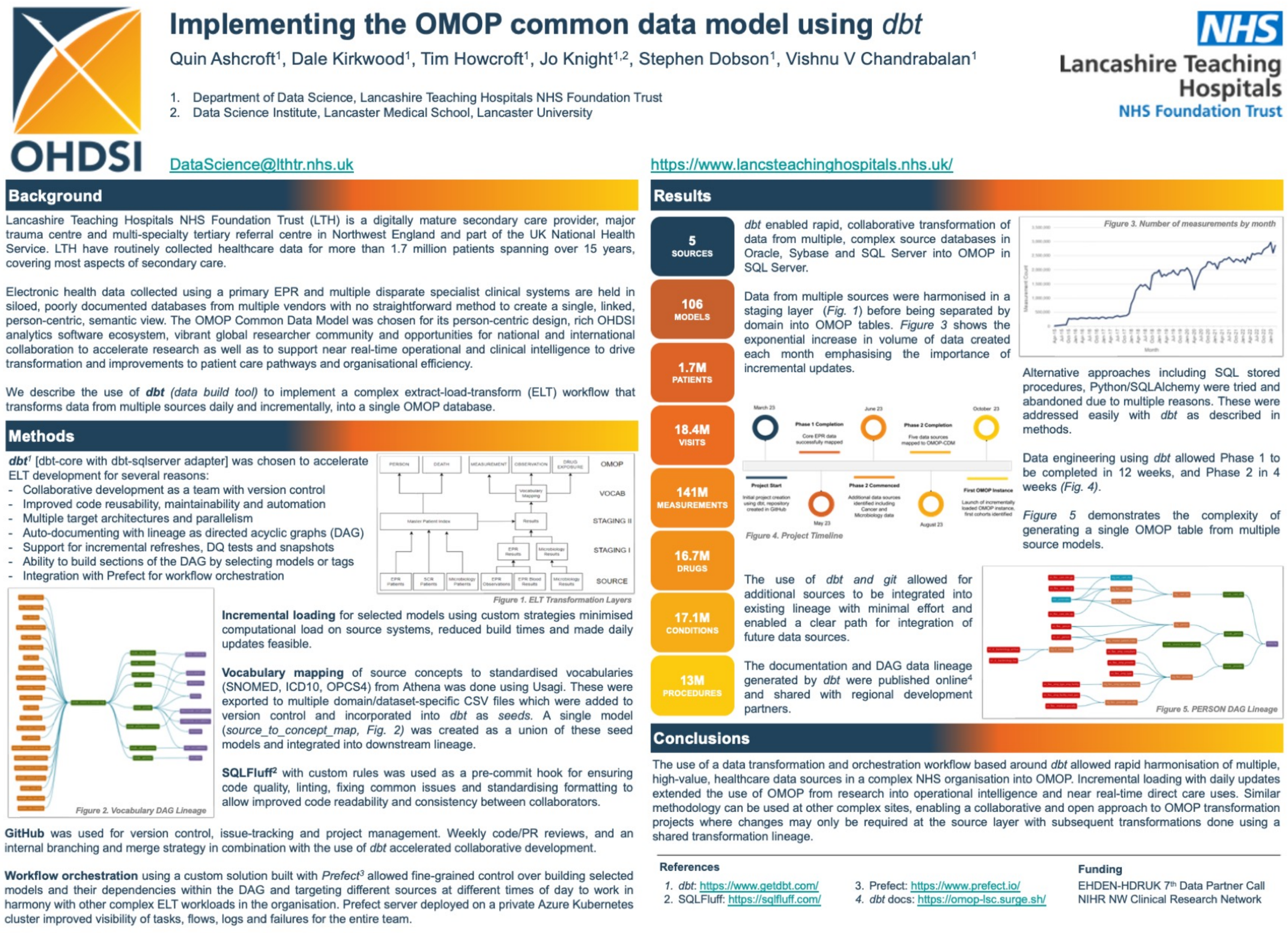
ohdsi-europe.org

#OHDSISocialShowcase This Week

MONDAY

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[1], Dale Kirkwood[1], Tim Howcroft[1], Jo Knight[1,2], Stephen Dobson[1], Vishnu V Chandrabalan[1]

1. Department of Data Science, Lancashire Teaching Hospitals NHS Foundation Trust
2. Data Science Institute, Lancaster Medical School, Lancaster University

DataScience@lthtr.nhs.uk

https://www.lancsteachinghospitals.nhs.uk/

OHDSI

NHS Lancashire Teaching Hospitals NHS Foundation Trust

Background

Lancashire Teaching Hospitals NHS Foundation Trust (LTH) is a digitally mature secondary care provider, major trauma centre and multi-specialty tertiary referral centre 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 siloed, 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 OHDSI analytics software ecosystem, vibrant global researcher 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

dbt[1] [dbt-core with dbt-sqlserver adapter] was chosen to accelerate ELT development for several reasons:

- Collaborative development as a team with version control
- Improved code reusability, maintainability and automation
- Multiple target architectures and parallelism
- Auto-documenting with lineage as directed acyclic graphs (DAG)
- Support for incremental refreshes, DQ tests and snapshots
- Ability to build sections of the DAG by selecting models or tags
- Integration with Prefect for workflow orchestration

Figure 1. ELT Transformation Layers

Figure 2. Vocabulary DAG Lineage

Incremental loading for selected models 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, OPCS4) from Athena was done using Usagi. These were exported to multiple domain/dataset-specific CSV files which were added to version control and incorporated into *dbt* as *seeds*. A single model (*source_to_concept_map, Fig. 2)* was created as a union of these seed models and integrated into downstream lineage.

SQLFluff[2] with custom rules was used as a pre-commit hook for ensuring code quality, linting, fixing common issues and standardising formatting to allow improved code readability and consistency between collaborators.

GitHub was used for version control, issue-tracking and project management. Weekly code/PR reviews, and an internal branching and merge strategy in combination with the use of *dbt* accelerated collaborative development.

Workflow orchestration using a custom solution built with *Prefect*[3] allowed fine-grained control over building selected models and their dependencies within the DAG and targeting different sources at different times of day to work in harmony with other complex ELT workloads in the organisation. Prefect server deployed on a private Azure Kubernetes cluster improved visibility of tasks, flows, logs and failures for the entire team.

Results

5 SOURCES · 106 MODELS · 1.7M PATIENTS · 18.4M VISITS · 141M MEASUREMENTS · 16.7M DRUGS · 17.1M CONDITIONS · 13M PROCEDURES

dbt enabled rapid, collaborative transformation of data from multiple, complex source databases 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 exponential increase in volume of data created each month emphasising the importance of incremental updates.

Figure 3. Number of measurements by month

Figure 4. Project Timeline

Alternative approaches including SQL stored procedures, Python/SQLAlchemy were tried and abandoned due to multiple reasons. These were addressed easily with *dbt* as described in methods.

Data engineering using *dbt* allowed Phase 1 to be completed in 12 weeks, and Phase 2 in 4 weeks *(Fig. 4)*.

Figure 5 demonstrates the complexity of generating a single OMOP table from multiple source models.

The use of *dbt and git* allowed for additional sources to be integrated into existing lineage 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[4] and shared with regional development partners.

Figure 5. PERSON DAG Lineage

Conclusions

The use of a data transformation and orchestration workflow based around *dbt* allowed rapid harmonisation of multiple, high-value, healthcare data sources in a complex NHS organisation into OMOP. Incremental loading with daily updates extended the use of OMOP from research into operational intelligence and near real-time direct care uses. Similar methodology can be used at other complex sites, enabling a collaborative and open approach to OMOP transformation projects where changes may only be required at the source layer with subsequent transformations done using a shared transformation lineage.

References

1. *dbt*: https://www.getdbt.com/
2. SQLFluff: https://sqlfluff.com/
3. Prefect: https://www.prefect.io/
4. *dbt* docs: https://omop-lsc.surge.sh/

Funding

EHDEN-HDRUK 7th Data Partner Call
NIHR NW Clinical Research Network

#OHDSISocialShowcase This Week

TUESDAY

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)

Mining Data Outside the Box: Internet as a New Source for Common Data Model

< Min-Gyu Kim MD>[1,2], < Min Ho An, MD >[1,2], <GyuBeom Hwang MD>1,2, <Rae Woong Park, MD, Ph.D.>[1]
1Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea
2Department of Medical Sciences, Graduate School of Ajou University, Suwon, Republic of Korea

Background

While the Observational Medical Outcomes Partnership(OMOP) Common Data Model (CDM) standardizes data acquired in healthcare settings, EHR 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.

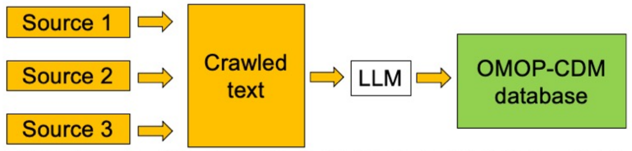


Fig.1) Description of the workflow of a generalized system that can mine medical data from various sources.

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.

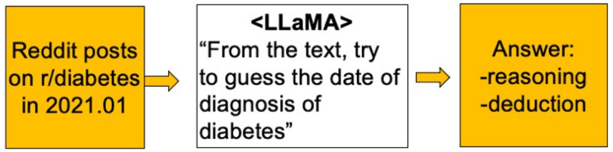


Fig.2) Diagram of the flow of this study. The model was asked to make a logical deduction about the date of diagnosis.

Results

Among the 200 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 found to be 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.

	Correct	Wrong	%
Fact	196	4	98.0
Logic	35	14	71.4
Math	27	7	79.4

Table.1) Three domains of accuracy. While there were few factual inconsistencies, logical integrity was less optimal.

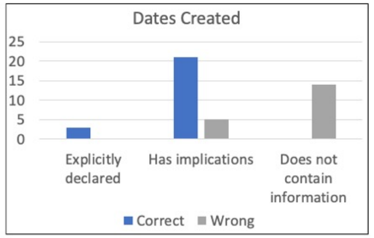


Fig.3) Dates extracted from 3 categories of posts. In cases where the date was explicitly mentioned, all dates were correctly extracted. When the post had indirect implications about the date, more than 80% were correctly extracted.

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 EHR alike. With the development of additional modules to assist LLMs, the internet may become a new source of medical data to feed OMOP-CDM.

Acknowledgement

•This research was funded 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: HR16C0001).
•This research was supported by a Government-wide R&D Fund project for infectious disease research (GFID), Republic of Korea (grant number: HG22C0024).

Contact: manjmin6@gmail.com

#OHDSISocialShowcase This Week

WEDNESDAY

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)

OHDSI

Forecasting daily incidence of Respiratory Symptoms: A Comparative Study on Time Series Models using OMOP-CDM in South Korea

<Min Ho An, MD>[1,2], <Min-Gyu Kim, MD>[1,2], <GyuBeom Hwang, MD>[1,2], <ByungJin Choi, MD>[1,2], <Rae Woong Park, MD, Ph.D>[1]
1 Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea
2 Department of Medical Sciences, Graduate School of Ajou University, Suwon, Republic of Korea

Background

- With the outbreak of the COVID-19 pandemic, the significance of infectious disease surveillance and upsurge prediction has been emphasized. Several reports associated with prediction of respiratory infectious disease including COVID-19 is published.
- Respiratory infectious diseases like COVID-19 can disseminate rapidly, given the impossibility of restricting respiratory activities. To monitor disease spread, four distinct hospitals in South Korea recently began to collaborate to collect data using the Observational Medical Outcomes Partnership - Common Data Model (OMOP-CDM) under the project named PHAROS (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, dyspnea, or cough at Ajou University Hospital in South Korea were defined as respiratory symptom related visit.
- 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 to train and test data, first allocating 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 v3.7.

Conclusions

- In the task of predicting daily counts of respiratory symptoms in South Korea, the ARIMA and Prophet models mostly presented 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 data'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 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: HR16C0001).
- This research was supported by a Government-wide R&D Fund project for infectious disease research (GFID), Republic of Korea (grant number: HG22C0024).

Results

- Table 1 reveals a marked decrease in visits since 2020. Both ARIMA (Fig. 1) and Prophet (Fig. 2) forecasts demonstrate similar outcomes, with most forecasted values lying within the 95% confidence interval for both models.
- Yet, the ARIMA model reported lower Mean Absolute Error (MAE) [2.66 vs 2.87] and Root Mean Squared Error (RMSE) [3.34 vs 13.10] than the Prophet model for test data (Fig 1-2).
- Intriguingly, the Prophet model better reflected the variance of the observed values than ARIMA, which primarily illustrated the downtrend with minimal variance.

Table 1. Summary of number of visits and respiratory symptoms in each year

	number of visits	fever	cough	dyspnea
2018	5523(29.3%)	3315(29.7%)	422(42.7%)	1786(26.7%)
2019	5563(29.5%)	3492(31.3%)	301(30.5%)	1770(26.5%)
2020	3775(20.0%)	2254(20.2%)	125(12.7%)	1396(20.9%)
2021	3978(21.1%)	2107(18.9%)	140(14.2%)	1731(25.9%)
Total	18839	11168	988	6683

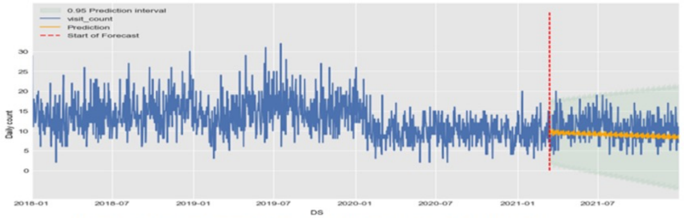


Figure 1. Daily count forecast using the ARIMA model. MAE: 2.66, RMSE: 3.34

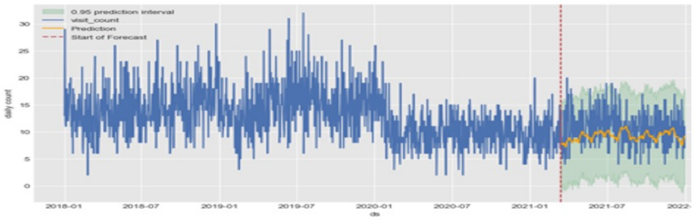


Figure 2. Daily count forecast using the prophet model. MAE: 2.87, RMSE: 13.10

#OHDSISocialShowcase This Week

THURSDAY

Observational Research in Dentistry: A Scoping Review

(Robert Koski, Danielle Boyce, Brock Johnson, Adam Bouras, Swetha Kiranmayi Jakkuva)

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-ScR 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 discusses a common data model or standardized terminology
- Discuss the implications, challenges, and or attempts to conduct observational research in dentistry
- Discuss implementation of a common data model in a given healthcare setting or specialty

Exclusion criteria:

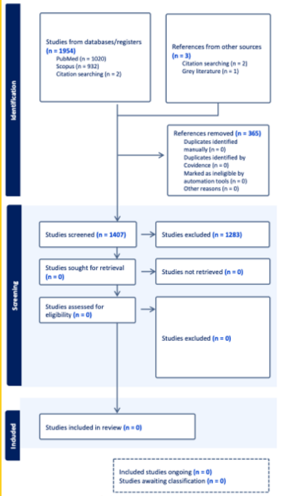
- Article published before 2010
- Article is not related to observational research
- Article does not pertain to the process of conducting observational research with health data
- Letters to the editor, editorials, critical reviews

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.

Take a picture to **download** the **full paper**

RESULTS



KEY FINDINGS

Current State

- Observational research capabilities in dentistry are nascent but growing
- The first observational research study in dentistry published in 2022
- National Dental Practice Based Research Network study used structured field data

Challenges

- No widely adopted standard terminology
- Terminologies that do exist have quality issues
- Dental records have quality issues that complicate research efforts
- OMOP-CDM lacks adequate coverage of dental concepts
- Variable reporting of diagnoses and findings among providers

Robert Koski, Danielle Boyce, Brock Johnson, Adam Bouras, Swetha Kiranmayi Jakkuva

Johns Hopkins School of Medicine
GVW Global Value Web
Tritonis inc
OHDSI

#OHDSISocialShowcase This Week

FRIDAY

Integration of Scalable Natural Language Processing to the Atlas Cohort Building Workflow

(Pavan Parimi, Selvin Soby, Pavel Goriacko, Chandra has Nelapatla, Boudewijn Aasman, Manuel Wahle, Reetam Nath, Parsa Mirhaji)

Integration of Scalable Natural Language Processing to the Atlas Cohort Building Workflow

EINSTEIN Albert Einstein College of Medicine

Parsa Mirhaji, Selvin Soby, Pavan Parimi, Chandra Nelapatla, Manuel Wahle, Boudewijn Aasman, Reetam Nath, Pavel Goriacko

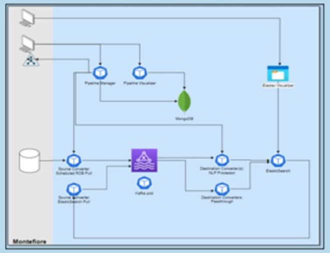
Montefiore

BACKGROUND

Einstein Atlas facilitates the design and execution of computable phenotyping and cohort-based analysis on standardized, patient-level, observational data. There are several methodologies utilized by the community to build cohorts from data derived from clinical text on OHDSI's OMOP datasets.[1,2,3] This project describes a mechanism for researchers to build advanced cohorts using discrete data from the OMOP CDM as well as concepts derived from clinical text using an NLP engine built on cTAKES and Elastic Search.

Clinical Text Analysis and Knowledge Extraction System (cTAKES) is an open-source natural language processing system for information extraction from EHR's clinical free-text. The system is based on Unstructured Information Management Architecture (UMIA) framework and the OpenNLP toolkit. The components are specifically trained for the clinical domain, based on Java and can be used to identify and extract entities specific entities, relationships between those entities, part-of-speech tagging, and dependency parsing.[2]

Figure 1: Architecture



METHODS

A: Data pre-processing: 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 labs by data informaticists.

B: Dictionary integration and ElasticSearch storage: The cTAKES software is configured, including the integration of the UMLS and OMOP dictionaries, relation-extraction, negation and context extraction, and any other required components. The output from cTAKES and other analytic engines, along with the related metadata, is serialized as 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 uniform text deidentification process that uses a large language model to tag protected information in clinical text. The text body, metadata, and annotations generated by the analytic pipelines are then placed into the Elastic Search database. Finally, necessary indexing is implemented to facilitate efficient retrieval of the processed 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 versatile cohort-based query engine was developed to enable submitting complex pattern search queries or terminology-based queries to JSON-LD and free text components. This engine allows query notes within an Atlas generated cohort and to integrate the results to research baskets and/or Atlas cohorts that could be further characterized, shared, or analytically used inside the Atlas framework. A specialized visualizer, browser was developed to assist with reviewing and navigating clinical documents as well as extracted metadata and query results for validation.

E: IRB Audit Trail: All NLP queries are stored and tracked for IRB auditing and/or sharing with other users.

Figure 2: Text-based Query Builder

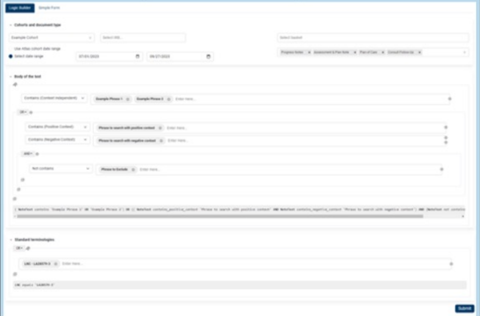


Figure 3: Text-based Query Builder

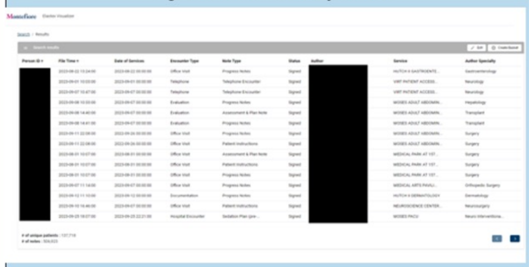


Figure 4: Example clinical text with annotations, mappings and deidentification

RESULTS

As of October 2023, over 134 million notes have been ingested and annotated by the NLP platform. Each note's annotations are made available through the scalable infrastructure. Four cohort projects which required the use of clinical text were attempted as a proof of concept using the platform. Each cohort's clinical text component was successfully completed using the NLP engine. Result lists for the queries are retrieved in microseconds, and the just-in-time deidentification of the note text is currently completed at an average rate of around 2 seconds. The deidentification is currently being optimized to distinguish between names of people and concept/medical terminology names. Research users have been able to successfully create complex cohorts using variables obtained from the clinical text.

CONCLUSION

The NLP engine that we developed and integrated to Atlas has provided our users the ability to build advanced cohorts using discrete data from the OMOP-CDM as well as concepts derived from the clinical text. The cTAKES and Elastic Search backend has been successfully implemented to provide extremely quick annotations and retrieval of text, most of which is completed in microseconds for real-time, highly scalable searches. The addition of LLM enabled deidentification is processed just-in-time for the research user which enables them to successfully create cohorts using highly complex queries pursuant to their needs without any preordination. IRB linking with systems such as IRIS and BRAINY allows for research users to view de-identified notes while maintaining an audit trail to protect PHI. The results of this project demonstrate the potential of integrating NLP engines to Atlas's cohort building capabilities.

REFERENCES

1. Yuan, C., Ryan, P. B., Ta, C. N., Guo, Y., Li, Z., Hardin, J. R., Makadia, R., Jin, P., Shang, N., Kang, T., & Weng, C. (2019). Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association, 26*(4), 294–305. https://doi.org/10.1093/jamia/ocy178
2. Tu, S. W., Peleg, M., Carini, S., Bobak, M. T., Ross, J. M., Rubin, D. L., & Sim, I. (2011). A practical method for transforming free-text eligibility criteria into computable criteria. *Journal of Biomedical Informatics, 44*(2), 239–250. https://doi.org/10.1016/j.jbi.2010.09.007
3. Weng, C., Wu, X., Luo, Z., Boland, M. R., Theodoratos, D., & Johnson, S. M. (2011). EliXR: an approach to eligibility criteria extraction and representation. *Journal of the American Medical Informatics Association, 18*(Supplement 1), i116–i124. https://doi.org/10.1136/amiajnl-2011-000321
4. Savova, Guergana; Masanz, James; Ogren, Philip; Zheng, Jiaping; Sohn, Sunghwan; Kipper-Schuler, Karin and Chute, Christopher. 2010. Mayo Clinic Clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. JAMIA 2010;17:507-513 doi:10.1136/jamia.2009.001560

Opening: Biomedical Informatics Data Scientist at Stanford

Stanford MEDICINE
Health Care

Who We Are | What We Offer | How We Hire | Career Areas | Search

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 •

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.

Postdoc/Senior Data Analyst Opening at WashU

PI: Linying Zhang, PhD

The Zhang Lab at Washington University School of Medicine in St. Louis has **one postdoct/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: 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

Apply

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.

About Us

GILEAD

Gilead Sciences, Inc. is a biopharmaceutical company that has pursued and achieved breakthroughs in medicine for more than three decades, with the goal of creating a healthier world for all people. The company is committed to advancing innovative medicines to prevent and treat life-threatening diseases, including HIV, viral hepatitis and cancer. Gilead operates in more than 35 countries worldwide, with headquarters in Foster City, California.

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

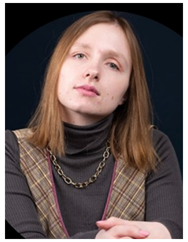
April 9: Vocabulary Techniques for ETL

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 Week 2 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