



ATLAS Deepdive: Characterization, Incidence, and Treatment Pathways

OHDSI Community Call
June 24, 2025 • 11 am ET





Upcoming Community Calls

Date	Topic
June 24	ATLAS Deepdive: Characterization, Incidence and Pathways
July 1	ATLAS Deepdive: Technical and Administrative Capabilities
July 8	No Meeting – Europe Symposium
July 15	Europe Symposium Review
July 22	OMOP/OHDSI Research Spotlight
July 29	Asia-Pacific Regional Updates
Aug. 5	No Meeting
Aug. 12	Newcomer Introductions



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	ATLAS
Wednesday	9 am	Oncology Outreach/Research Subgroup
Wednesday	10 am	Surgery and Perioperative Medicine
Wednesday	10 pm	Women of OHDSI
Wednesday	11 am	Common Data Model
Wednesday	12 pm	Latin America
Wednesday	7 pm	Medical Imaging
Thursday	9:30 am	Network Data Quality
Friday	9 am	Phenotype Development and Evaluation
Friday	10 am	GIS – Geographic Information System
Friday	10 am	Transplant
Friday	11 am	Clinical Trials
Friday	11:30 am	Steering
Monday	10 am	Healthcare Systems Interest Group

Is Semaglutide Associated with Yet Another Blinding Eye Disease?

JAMA Ophthalmology | **Original Investigation**

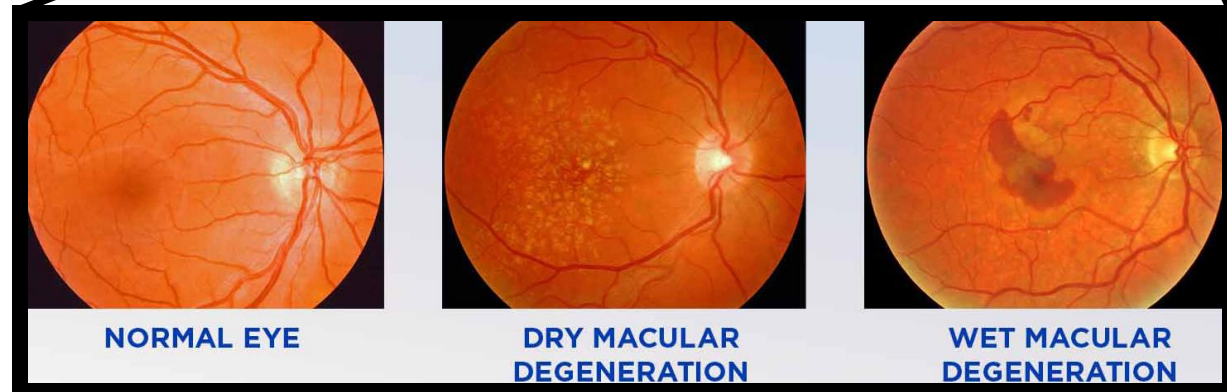
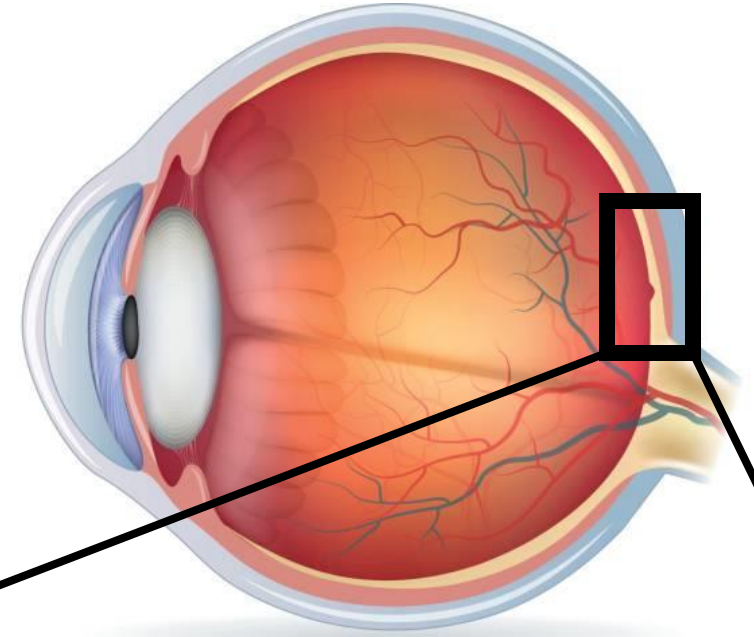
Glucagon-Like Peptide-1 Receptor Agonists and Risk of Neovascular Age-Related Macular Degeneration

Reut Shor, MD; Andrew Mihalache, MD(C); Atefeh Noori, PhD; Renana Shor, MD; Radha P. Kohly, MD, PhD; Marko M. Popovic, MD, MPH; Rajeev H. Muni, MD, MSc

Hazard Ratio of NVAMD 2.21 (95% CI 1.65 – 2.96)

Linked claims + EHR data (Ontario Health Insurance Plan)

46,334 adults with diabetes exposed to GLP1-RA (>6mo) compared to 92,668 unexposed to GLP1-RA

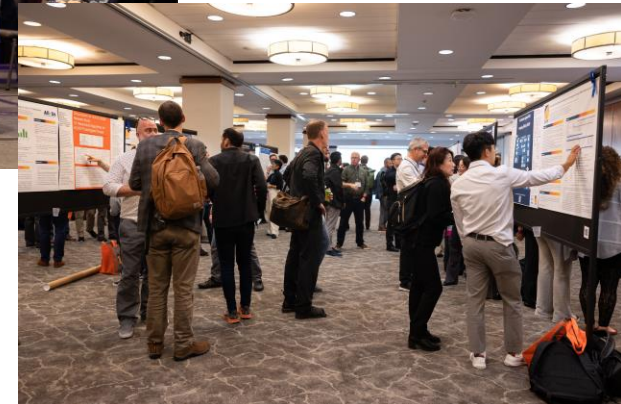




ONE Week Remaining

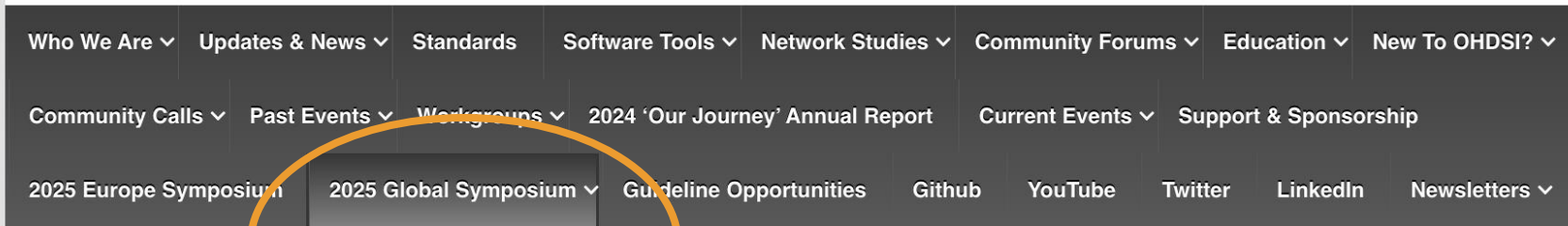
The submission deadline for the 2025 Global Symposium Collaborator Showcase is **July 1 (8 pm ET)**.

More information about the collaborator showcase, including links to the submission form and poster templates, can be found on the #OHDSI2025 homepage.





Global Symposium: Oct. 7-9



2025 OHDSI Global Symposium

Oct. 7-9 • New Brunswick, N.J. • Hyatt Regency Hotel

There is nothing quite like the OHDSI Global Symposium, which welcomes hundreds of collaborators around the world who believe in the shared mission of improving health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care. We can't wait to return for our biggest event of the year this October in New Brunswick, N.J.



2025 Africa Symposium

The 2025 OHDSI Africa Symposium will be held Nov. 10-12 in Kampala, Uganda.

The abstract submission deadline will be August 25.

OHDSI AFRICA Symposium

Hosted by:
Joint Clinical Research Centre
KAMPALA, UGANDA

Abstract Submission Deadline
August 25, 2025

Save The Date
NOV 10th - 12th 2025

About the Symposium:
Discover how interoperable healthcare data and federated networks can unlock real-world insights while safeguarding patient privacy. This innovative approach - bringing analyses to the data - empowers large-scale, global collaboration without compromising data ownership. A key highlight will be on advancing data-driven healthcare solutions, with a special focus on HIV.

Organised by JCRC
in collaboration with: OHDSI

www.jcrc.org.ug **JCRC**

OHDSI Uganda



2025 Africa Symposium

Draft Agenda (Very Preliminary)

Monday, Nov 10: Tutorials

Fundamentals: Standardized Vocabularies

Fundamentals: The OMOP CDM and ETL Process

Fundamentals: Data Quality & ATLAS installation

Fundamental Building Blocks to Evidence; Examples of Analytics

Individual Consultations



2025 Africa Symposium

Draft Agenda (Very Preliminary)

Tues, Nov 11 Morning Session (OHDSI only)

Welcome from JCRC

Uganda Minister of Health Informatics Division

Uganda Minister of Science, Technology and Innovation

History of OHDSI Africa

JCRC's Journey with OHDSI

OpenMRS to the OMOP CDM

Data Science Without Borders

Interoperability of Mental Health Data

Panel Discussion with morning speakers

Tues, Nov 11 Afternoon Session (Joint with HIV Conference)

JCRC Executive Director

Uganda Minister of Health

Frank Graziano Memorial Lecture on HIV

Highlights from Past Years' Int'l Conferences on HIV

Generating Reliable Evidence from RWD & Addressing HIV Evidence Gaps

Malawi HIV Data Lake

PEPFAR Update

Generating PEPFAR Statistics from OMOP'd data

Advances in Anti-retroviral Treatments

Treatment Pathways in HIV Therapy

Building Human Capacity: BRIDGE Training Grant



2025 Africa Symposium

Draft Agenda (Very Preliminary)

Wed, Nov 12 Morning Session (OHDSI only)

Speaker from the Africa CDC

Speaker from African Health Data Space

Standardizing terminology unique to African Context

Maternal Health in the Western Cape & Brazil

Harmonizing Mental Health data to the OMOP CDM

VODAN Antenatal Care Analysis using OHDSI Tools

Data Science Without Borders

HELINA Speaker

Lightening talks

Wed, Nov 12 Afternoon Session (Joint with HIV Conference)

Converting Household Demographic Survey Data to the OMOP CDM

Pediatric AIDS

African Population Cohort Consortium

Tuberculosis and HIV

Geospatial data representation

Mpox and Marburg

HIV Vaccine

DARWIN-EU

Book of OHDSI Translations into French, Portuguese, Arabic & Kishwari

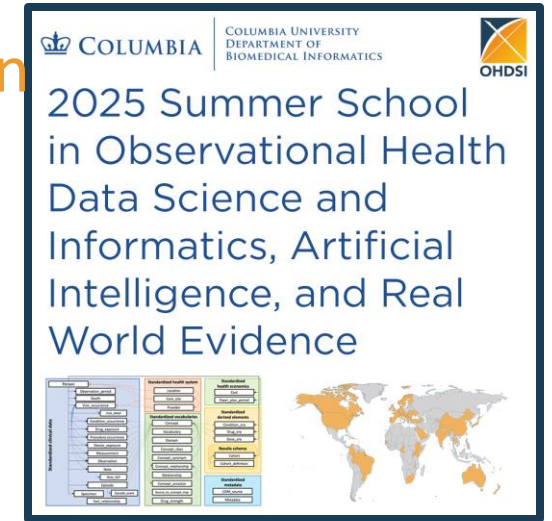
Panel Discussion



Columbia Summer School on OHDSI

Registration is open for the first ever Columbia Summer School on OHDSI, held July 14-18, 2025, at the Columbia University Department of Biomedical Informatics in New York City.

The Columbia Summer School in Observational Health Data Science and Informatics, Artificial Intelligence, and Real World Evidence (RWE) offers health professionals, researchers and industry practitioners the opportunity to gain familiarity and hands-on experience with real world data and generating real world evidence. Participants will learn about the different types of healthcare data captured during routine clinical care, including electronic health records and administrative records, and how these data can be standardized to the OMOP Common Data Model to enable distributed data network research.



Meet Our Faculty



George Hripcsak, MD MS
Vivian Beaumont Allen
Professor of Biomedical
Informatics



Patrick Ryan, PhD
Adjunct Assistant
Professor of Biomedical
Informatics



Anna Ostropolets, MD PhD
Adjunct Assistant
Professor of Biomedical
Informatics



Karthik Natarajan, PhD
Assistant Professor of
Biomedical Informatics



#OHDSISocialShowcase This Week

Monday

Electronic Frailty index and hazard of with MACE event in patients with Type 2 diabetes mellitus

(Da Eun Hyeon, Sujin Gan, Rae Woong Park)



Electronic Frailty index and hazard of with MACE event in patients with Type 2 diabetes mellitus

Daeun Hyeon¹, Sujin Gan¹, Rae Woong Park^{1,2}

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

- Type 2 diabetes mellitus (T2DM) significantly increases the risk of cardiovascular disease (CVD), particularly in older individuals with frailty and multimorbidity.
- Traditional frailty assessment tools are often complex and time-consuming, highlighting the need for EMR-based frailty indices to identify at-risk patients more efficiently.
- This study aims to investigate the association between frailty, as measured through electronic medical records (EMRs), and major adverse cardiovascular events (MACE) in T2DM patients.

Methods

1. Data preparation

- Observational medical outcomes partnership common data model (OMOP-CDM) database at Ajou University School of Medicine (AUSOM)

• Inclusion criteria

- 1) 40 years and older
- 2) Diagnosed with type 2 diabetes mellitus (T2DM)
- 3) No history of major adverse cardiovascular events (MACE) ; myocardial infarction, cardiovascular disorders, acute ischemic heart disease, chronic ischemic heart disease and acute myocardial infarction.

2. Outcome

- Occurrence of MACE

3. Sensitivity analysis

- Dividing the participants into two age groups : 65 years and younger and 66 years and older.

3. Frailty index calculation

- Electronic medical record (EMR) data was used to calculate the Electronic Frailty Index (eFI).
- The eFI was calculated by summing binarized variables, resulting in a score ranging from 0 to 1.
- This score was divided by the number of variables per patient, excluding missing values.
- The maximum value of eFI was divided into thirds, stratifying patients into three groups based on each interval.
- Patients were categorized as normal, pre-frailty, or frailty based on their FI score.

4. Statistical Analysis

- Cox proportional hazards regression model
- Kaplan-Meier survival curves
- Log-rank test

Conclusions

- This study shows an association between increasing eFI and the occurrence of MACE in patients with T2DM aged 40 years or older.
- The eFI used in this study has the advantage of not requiring separate frailty testing, and it showed the feasibility of using eFI in OMOP-CDM to screen for CVD risk groups in patients with T2DM.

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

- The risk of MACE was significantly higher in the frailty group compared to the normal group (Hazard Ratio [HR]: 1.68, 95% Confidence Interval [CI]: 1.38-2.04; P < 0.05) and in the pre-frailty group compared to the normal group (HR 1.44 (1.35-1.55); P < 0.05).

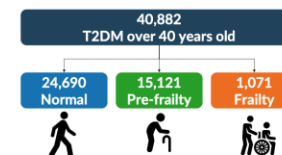


Figure 1. Classification of T2DM patients based on electronic Frailty Index

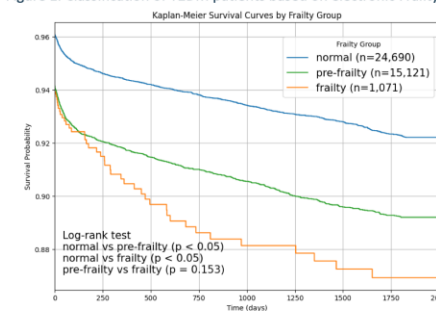


Figure 2. Survival probability curve for normal, pre-frailty, and frailty groups

Age	Frailty category	N	Hazard ratio (95% CI)
Below 65	Normal	15,714	1.0 (ref)
	Pre-Frailty	8,028	1.26 (1.13-1.40)
	Frailty	505	1.28 (0.90-1.83)
Over 65	Normal	8,976	1.0 (ref)
	Pre-Frailty	7,093	1.46 (1.33-1.61)
	Frailty	566	1.70 (1.34-2.16)

Table 1. Hazard ratio for MACE in Subgroup



@OHDSI

www.ohdsi.org

#JoinTheJourney



ohdsi



#OHDSISocialShowcase This Week

Tuesday

An Explorative Study about the Latent Space of Clinical Foundation Models Based on a Common Data Model Database

(Min-Gyu Kim, Dong Yun Lee, Jinyang Kim, Joon-Kyung Seong, Rae Woong Park)



An Explorative Study about the Latent Space of Clinical Foundation Models Based on a Common Data Model Database

Min-Gyu Kim^{1,2}, Dong Yun Lee^{1,2}, Jin Yang Kim³, Rae Woong Park^{1,2}, Joon-Kyung Seong^{3,4}
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
³Department of Artificial Intelligence, Korea University, Seoul, South Korea
⁴School of Biomedical Engineering, Korea University, Seoul, South Korea

Background

Recently, there have been researches about clinical foundation models (FMs), which have shown advantages over traditional prediction model. While metrics like F1 score can explain the performance of a model objectively, they are usually inadequate for understanding the internal structure of the model. Also, methods to train such models are still limited to analogies from the language domain. There are many methods available that enable model understanding, such as visualizing self-attention of each layer or dimension reduction in the latent space. In this study, we aim to understand how we should train clinical foundation models by first training a model using our own data based on OMOP-CDM and visualizing the latent space of the trained model.

Methods

We trained a transformer model based on the bidirectional transformer (BERT) architecture, using data from Ajou university hospital standardized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Patient records were first translated into a time series format. Additional information such as patient age and gender were prepended to the input series as separate tokens. To provide a better understanding about the domains defined by OMOP-CDM, each token was added to the embedding about its domain, i.e. condition, drug, measurement. The model was trained using masked language modeling. 15% of the tokens were randomly masked and the model predicted the original tokens. 1% of the total training data was randomly selected, and the CLS tokens of the sample were calculated. The tokens were then reduced to seven dimensions using Uniform Manifold Approximation (UMAP) and clustered with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). The result was visualized using t-distributed stochastic neighbor embedding (t-SNE) by reducing to a 2-dimensional plane. The resulting visualization was inspected, and cluster formation was manually evaluated using Term Frequency-Inverse Document Frequency (TF-IDF).

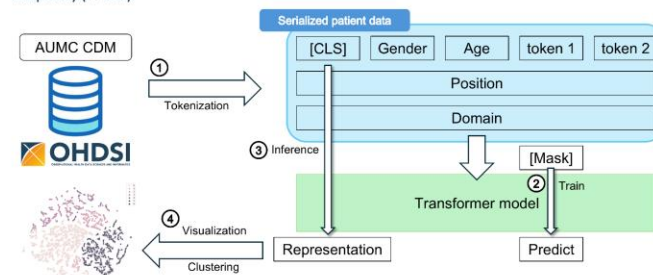


Figure 1. Study flow. First, the OMOP-CDM in Ajou university medical center was transformed into serial data according to each patient, including basic patient information such as gender and age. The data was then fed through a transformer model.

Contact: manjmin6@gmail.com / Min-Gyu Kim

Results

Training loss converged and the model with the least validation error was selected. The clusters were not immediately recognizable with the IDs only, but some was specific enough to make weak assumptions about the cluster. For example, cluster 5 had measurements related to health screening. The visualization of clusters using representative tokens showed better results in cluster membership. While some tokens representing a cluster was not present for most of the patient data within that cluster, certain tokens clearly showed patterns of grouping (Figure 2), closely resembling the distribution of the cluster.

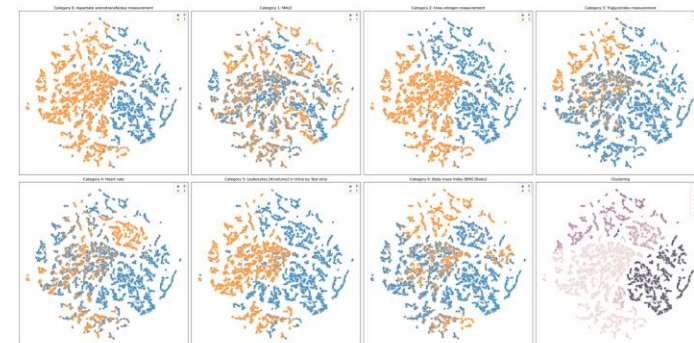


Figure 2. Top 1 representative token of each cluster visualized. (Blue) Patients without representative concept ID of cluster 0 to 6. (Orange) Patients with representative concept ID of cluster 0 to 6.

Conclusions

In this study, we trained a BERT-based clinical foundation model using data from electronic health record converted to OMOP-CDM. The latent space was visualized using dimension reduction techniques and clusters with explainable characteristics were found in some cases. A better optimized approach with different architectures or training method may lead to a better intuitive understanding about the data contained using 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) and this research was supported by a Government-wide R&D Fund project for infectious disease research (GFID), Republic of Korea (grant number: HG22C0024, KH124685).



#OHDSISocialShowcase This Week

Wednesday

Causal Learning with Large-Scale Propensity Scores to Predict Treatment Outcomes: A Study of Arrhythmia in Adolescents with Attention-deficit/hyperactivity disorder

(Junhyuk Chang, Dong Yun Lee, Rae Woong Park)



Causal Learning with Large-Scale Propensity Scores to Predict Treatment Outcomes : A Study of Arrhythmia in Adolescents with Attention-deficit/hyperactivity disorder

Junhyuk Chang, PharmD¹, Dong Yun Lee, MD², Rae Woong Park, MD, Ph.D.^{1,2}

¹Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, South Korea

²Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea



Background

- Adolescents with ADHD and comorbid depression often receive methylphenidate (MPH) and selective serotonin reuptake inhibitors (SSRIs)
- Concurrent use of MPH and SSRIs may increase cardiovascular risks, including arrhythmia
- The causal machine learning method is able to estimate treatment effects on individual patients by calculating average treatment effects.
- This study aims to analyze the treatment effect of concomitant administering SSRIs and MPH on arrhythmia occurrence with a causal forest model

Methods



Figure 1. Overall study framework

1. Data collection

- Database: Health Insurance Review and Assessment Service – Attention Deficit/Hyperactivity Disorder (HIRA-ADHD) database which contained ADHD patient data from nationwide claims data
- HIRA-ADHD database was converted to OMOP-CDM
- Data was collected from Jan 1, 2016 to Dec 31, 2020

2. Cohort definition

- Target Cohort**
 - MPH-used patients with an ADHD diagnosis aged between 10 and 19
 - Patients with a depression record
 - Patients without other anti-ADHD agents and previous antidepressants

Outcome Cohort: Occurrence of arrhythmia

3. Data preprocessing

- Split: 70% for training / 30% for testing, ensuring the same outcome prevalence in both sets
- Extracted patient baseline covariates to employ a large-scale propensity score utilizing the FeatureExtraction
- Initial screening was conducted to exclude rare covariates by 10-fold cross-validation

4. Estimate average treatment effect

- Estimated the average treatment effect (ATE) using constructed causal forest model
- Using rank-ATE (RATE), we estimated treatment heterogeneity based on the quintiles of the test set divided according to CATEs
- We compared the top 5 variables based on variable importance from the causal forest model to identify characteristics of high and low CATE groups

Contact: contact@ohdsi.org

Results

- Among the total of 11,163 MPH-used patients, 7,873 patients were prescribed SSRIs and 58 patients had occurrences of arrhythmia

- Figure 2 shows the ATEs of the quantile groups in increasing order, with values of -0.5, -0.1, 0.1, 0.1, and 0.4
- Among ATE of quantile groups, the ATE of the Q5 group is statistically significant (95% CI: 0.1-0.8).

- The estimated RATE was 0.008 (95% CI: 0.002-0.015), which confirmed the heterogeneity between quantile groups

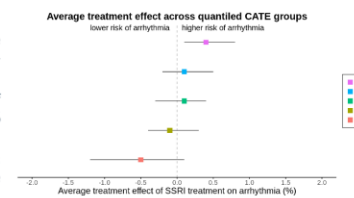


Figure 2. Average treatment effect of quantile groups

- Figure 3 represents the density of top 5 baseline covariates between high and low CATE groups

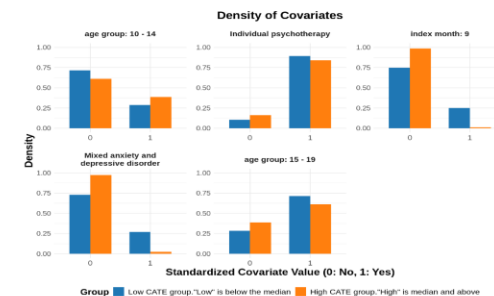


Figure 3. Density of top 5 covariates

Conclusions

- This study suggests that while SSRI treatment did not significantly affect arrhythmia
- Individualized treatment rule accounting for this heterogeneity could modify guidelines for concurrent use of MPH and SSRIs

Acknowledgements

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) and this research was supported by a Government-wide R&D Fund project for infectious disease research (GFID), Republic of Korea (grant number: HG22C0024, KH124685).



@OHDSI

www.ohdsi.org

#JoinTheJourney



ohdsi



#OHDSISocialShowcase This Week

Thursday

Leveraging Large Language Model for Populating OMOP Oncology CDM from the EHR: Feasibility Study

(Subin Kim, Jeong Eun Choi, Chang Jun Ko, Seng Chan You)

Leveraging Large Language Model for Populating OMOP Oncology CDM from the EHR

: Feasibility Study PRESENTER: Seng Chan You

INTRODUCTION

- The Oncology CDM Working Group developed the OMOP Oncology Extension to support the integration of cancer-specific information into the OMOP CDM.
- Despite these advancements, much of the cancer-data in EHR remains in unstructured formats, making it challenging to utilize and standardize.
- Generative large language models (LLMs) present a promising solution to these challenges, by leveraging the in-context learning capabilities of LLMs.
- In this study, we developed strategy to extract the cancer information from unstructured pathology and radiology reports of patients with colorectal cancer using state-of-the-art LLM.

- Among several candidate applications to validate feasibility, we focused on whether LLM-derived cancer data can be used to define cancer stage at diagnosis in accordance with updates to the AJCC staging system.

METHODS

Data sources

- We obtained unstructured pathology and radiology reports for patients diagnosed with colorectal cancer at Severance Hospital between 2010 and 2023.
- A random sample of 1,000 individuals was selected for inclusion in the study. We used 1,579 radiology and 2,632 pathology reports documented within 30 days before or 120 days after initial cancer diagnosis.

Prompt design

- We interacted with GPT-4o via zero-shot prompting through the OpenAI API. A total of 20 reports were sampled to develop prompts to extract cancer data (Table 1). All output was compiled into a JSON format.

Evaluation

- We classified the cancer stage at diagnosis using based on the 8th edition of the AJCC TNM staging system. We compared the LLM-derived cancer stage at diagnosis with the TNM values retrieved from the EHRs database.
- Additionally, we defined the cancer stage using both the 7th and 8th editions of the AJCC staging system and illustrated the changes in cancer stage, demonstrating the usefulness and flexibility of the LLM-derived cancer information.

Generative LLM can be used to populate Oncology CDM from the unstructured EHRs

Table 1. Oncologic data extracted from pathology and radiology reports

Pathology reports		Radiology reports	
Category	Descriptor	Category	Descriptor
Feature	Size	Lymph node	Metastasis site
	Histologic grade		Metastasis count
	Histologic type		BRAF mutation
	Location		KRAS mutation
	Procedure		Ki-67 index
	Tumor status		MLH1
	Depth of invasion		MSH2
	Lymphovascular invasion		MSH6
	Perineural invasion		Microsatellite instability
	Tumor budding		Mitotic count
Margin	Tumor deposits	Other	NRAS mutation
	Basal margin		PMS2
	Circumferential margin		
	In		
	Distal margin		
	Lateral margin		
Radiology reports	Proximal margin	Resection margin	
	Resection margin		
Radiology reports		Radiology reports	
Feature	Tumor location	Feature	Tumor status
	Size		Size

Figure 1. Overall performance of GPT-4o on classification of cancer stage

		AJCC staging from EHR									
		0	1	2A	2B	2C	3A	3B	3C	4A	4B
AJCC staging from LLM	0	58	1	0	0	0	0	1	3	0	0
	1	2	234	7	0	0	0	1	0	0	0
	2A	1	1	116	1	0	0	3	0	1	0
	2B	0	0	0	11	0	0	0	0	1	0
	2C	1	0	0	3	0	0	0	0	0	0
	3A	0	1	0	0	0	18	0	0	0	0
	3B	0	0	1	0	0	99	4	0	0	1
	3C	0	0	0	0	0	1	14	0	0	0
	4A	0	15	13	0	0	1	5	2	9	2
	4B	0	1	2	1	0	0	0	0	0	1
		IVC	0	0	5	1	1	0	2	4	2

Figure 2. Comparison of TNM staging according to the AJCC editions

		TNM staging from AJCC 7th edition							
		0	1	2A	2B	2C	3A	3B	3C
TNM staging from LLM	0	75	0	0	0	0	0	0	0
	1	0	245	0	0	0	0	0	0
	2A	0	0	103	0	0	0	0	0
	2B	0	0	0	13	0	0	0	0
	2C	0	0	0	0	4	0	0	0
	3A	0	0	0	0	0	19	0	0
	3B	0	0	0	0	0	0	108	0
	3C	0	0	0	0	0	0	15	0
	4A	0	0	0	0	0	0	0	46
	4B	0	0	0	0	0	0	0	0
		IVC	0	0	0	0	0	0	10

RESULTS

- A total of 4,211 pathology and radiology reports from 1,000 patients were analyzed.
- The agreement between LLM-derived AJCC stage and AJCC stage from structured EHRs is presented using confusion matrix in Figure 1. The overall accuracy of LLM-derived staging was 0.86. Cohen's Kappa was 0.82 (95% confidence interval [CI], 0.78-0.85).
- Figure 2 shows the comparison of TNM staging groups according to the AJCC 7th and 8th edition.

- A major difference between 7th and 8th edition is that the inclusion of new stage involving peritoneal metastasis (stage IVC).
- As a result, 19 patients, originally classified as stage IVC or IVB under the 7th edition, were reclassified as stage IVC.

CONCLUSION

- This is ongoing study. Generative LLMs demonstrate feasibility in automating the extraction of structured cancer information from unstructured EHRs.
- This approach has the potential to construct well-defined resources for future research, reducing the workload of human experts.

- By leveraging generative LLM, we will standardize the cancer-specific data from the EHR based on the OMOP Oncology Extension.

Subin Kim^{1,2}, Jeong Eun Choi^{1,2}, Chang Jun Ko¹, Seng Chan You^{1,2}

¹Dept. of Biomedical Systems Informatics, Yonsei University College of Medicine

²Institute for Innovation in Digital Health Care, Yonsei University

³Dept. of Health Informatics and Biostatistics, Graduate School of Public Health, Yonsei University



@OHDSI

www.ohdsi.org

#JoinTheJourney



ohdsi



#OHDSISocialShowcase This Week

Friday

Exploring Stroke and Cognitive Impaired Patients Using Apache Superset on OMOP OHDSI Dataset

(**Muhammad Solihuddin Muhtar**,
Phung Anh (Alex) Nguyen, Jason C. Hsu)

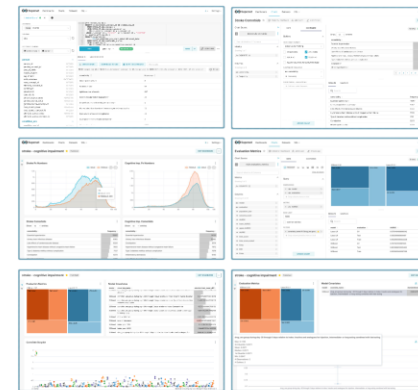


Apache Superset for rapid exploration tools for OHDSI's OMOP CDM

Exploring Stroke and Cognitive Impairment in the TMU-CRD OMOP CDM Dataset Using Apache Superset

Background: This study utilizes Apache Superset, an open-source BI tool as an alternative to Shiny App, to explore stroke and cognitive impairment data within the OMOP OHDSI framework, specifically leveraging the TMU Clinical Research Database (TMU-CRD), and to explore the OHDSI's PLP study result as well.

Methods: Exploratory data analysis conducted with Superset's intuitive drag-and-drop interface and SQL capabilities for dashboard creation. Data queries were built using OHDSI's QueryLibrary, facilitating analysis of comorbidities and demographic trends, extracted from TMU-CRD CDM database. Visualization for PLP Results were derived from OHDSI's PLP package result, including XGBoost and LASSO regression, revealing predictive insights from patient features.



Results: Some tabular and charts were presented interactively, and some adjustments could be easily facilitated through internal SQL Lab across multiple databases.

Conclusion: This approach showcases Apache Superset's flexibility and accessibility for exploring large-scale health datasets. Its BI capabilities empower researchers to visualize and interpret complex patterns in stroke and cognitive impairment, paving the way for further clinical insights.

Limitation: While Apache Superset excels in flexibility and ease of use, it faces challenges in replicating standard statistical and evaluation metrics curves (e.g., ROC or calibration curves) commonly produced in dedicated statistical software. These limitations may require external tools or programming to supplement Superset's functionality for advanced model evaluations.



Muhammad Solihuddin Muhtar¹, Nguyen Phung-Anh^{2,3,4,5} Jason C. Hsu^{1,3,4,5} Min-Huei Hsu^{1,2,3,5}

¹ International Ph.D. Program in Business and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan
² Graduate Institute of Data Science, College of Management, Taipei Medical University, Taipei, Taiwan
³ Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan
⁴ Research Center of Health Care Research Data Science, College of Management, Taipei Medical University, Taipei, Taiwan
⁵ Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei Medical University, Taipei, Taiwan



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?





June 24: ATLAS Deepdive

Characterization, Incidence, Treatment Pathways



Christopher Knoll

Director, Observational Health Data Analytics
Janssen Research and Development
ATLAS Workgroup Lead

Join us
throughout June to help
create the roadmap for
ATLAS!



**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-2025**