Current Practices in Estimation and Prediction

OHDSI Community Call
April 22, 2025 • 11 am ET

OHDSI
Observational Health Data Sciences and Informatics

Upcoming Community Calls

Date	Topic
Apr. 22	Current Practices in Estimation and Prediction
Apr. 29	DevCon 2025 Review
May 6	Evidence Synthesis
May 13	Maternal Health Fellowship Review
May 20	Guideline-Driven Evidence Study Review
May 27	Collaborator Showcase Brainstorm (Deadline is July 1)

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Daniel Kapitan, Femke Heddema, André Dekker, Melle Sieswerda, Bart-Jan Verhoeff, and Matt Berg** on the publication of **Data Interoperability in Context: The Importance of Open-Source Implementations When Choosing Open Standards** in the *Journal of Medical Internet Research.*

Viewpoint

Data Interoperability in Context: The Importance of Open-Source Implementations When Choosing Open Standards

Daniel Kapitan[1,2,3], DPhil; Femke Heddema[2], MSc; André Dekker[4], Prof Dr; Melle Sieswerda[5], MD, MSc; Bart-Jan Verhoeff[6], MD; Matt Berg[7], BA, MBA

[1]Eindhoven AI Systems Institute (EAISI), Eindhoven University of Technology, Eindhoven, The Netherlands
[2]PharmAccess Foundation, Amsterdam, The Netherlands
[3]Dutch Hospital Data, Utrecht, The Netherlands
[4]MAASTRO Clinic, Maastricht University Medical Centre, Maastricht University, Maastricht, The Netherlands
[5]Netherlands Comprehensive Cancer Organisation, Utrecht, The Netherlands
[6]Expertisecentrum Zorgalgoritmen, Utrecht, The Netherlands
[7]Ona, Burlington, VT, United States

Corresponding Author:
Daniel Kapitan, DPhil
Eindhoven AI Systems Institute (EAISI)
Eindhoven University of Technology
PO Box 513
Eindhoven, 5600 MB
The Netherlands
Phone: 31 624097295
Email: daniel@kapitan.net

Abstract

Following the proposal by Tsafnat et al (2024) to converge on three open health data standards, this viewpoint offers a critical reflection on their proposed alignment of openEHR, Fast Health Interoperability Resources (FHIR), and Observational Medical Outcomes Partnership (OMOP) as default data standards for clinical care and administration, data exchange, and longitudinal analysis, respectively. We argue that open standards are a necessary but not sufficient condition to achieve health data interoperability. The ecosystem of open-source software needs to be considered when choosing an appropriate standard for a given context. We discuss two specific contexts, namely standardization of (1) health data for federated learning, and (2) health data sharing in low- and middle-income countries. Specific design principles, practical considerations, and implementation choices for these two contexts are described, based on ongoing work in both areas. In the case of federated learning, we observe convergence toward OMOP and FHIR, where the two standards can effectively be used side-by-side given the availability of mediators between the two. In the case of health information exchanges in low and middle-income countries, we see a strong convergence toward FHIR as the primary standard. We propose practical guidelines for context-specific adaptation of open health data standards.

OHDSI Shoutouts!

Congratulations to the team of **Young Hwa Lee, Young June Choe, Yoon Sun Yoon, Ji Young Park, Yun-Kyung Kim, Hyung Joon Joo, Sujin Choi, Hyun Jung Kim and Lorenzo Bertizzolo** on the publication of **Predicting ICU Admission Risk in Children with Respiratory Syncytial Virus** in *Infectious Diseases and Therapy.*

Infect Dis Ther
https://doi.org/10.1007/s40121-025-01155-w

Check for updates

ORIGINAL RESEARCH

Predicting ICU Admission Risk in Children with Respiratory Syncytial Virus

Young Hwa Lee · Young June Choe · Yoon Sun Yoon · Ji Young Park · Yun-Kyung Kim · Hyung Joon Joo · Sujin Choi · Hyun Jung Kim · Lorenzo Bertizzolo

Received: February 26, 2025 / Accepted: April 2, 2025

ABSTRACT

***Introduction*:** Respiratory syncytial virus (RSV) is a common infection in young children and a frequent cause of hospitalization. In some cases, RSV can lead to severe lower respiratory tract illness requiring admission to the intensive care unit (ICU). Here, we explore risk factors for RSV-related ICU admission in children.
***Methods*:** We conducted a retrospective study using Electronic Medical Record (EMR) data transformed into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) from three tertiary care centers in Korea between 2008 and 2022. We identified 1529 children hospitalized with RSV according to the CDM and examined risk factors for ICU admission in this population.
***Results*:** Of 33,674 children aged 0–9 years who tested for RSV, 1529 (4.5%) were positive. The highest proportion of RSV-positive children were less than 10 months old. The ICU admission rate among RSV-positive children was 1.8% (29/1529), and the highest ICU admission rate occurred in children aged 0–5 months (4.4%). In a multivariable logistic regression model, we found that the

Prior Presentation: Part of this study was presented at the RSVVW'24 Conference, 13th-16th February 2024, in Mumbai, India.

OHDSI Shoutouts!

Congratulations to the team of **Jenna Reps, Peter Rijnbeek and Patrick Ryan** on the publication of **Can we develop real-world prognostic models using observational healthcare data? Large-scale experiment to investigate model sensitivity to database and phenotypes** in *Diagnostic and Prognostic Research.*

Reps *et al.*
Diagnostic and Prognostic Research (2025) 9:10
https://doi.org/10.1186/s41512-025-00191-x

Diagnostic and Prognostic Research

RESEARCH **Open Access**

Check for updates

Can we develop real-world prognostic models using observational healthcare data? Large-scale experiment to investigate model sensitivity to database and phenotypes

Jenna M. Reps[1,2*], Peter R. Rijnbeek[2] and Patrick B. Ryan[1]

Abstract

Background Large observational healthcare databases are frequently used to develop models to be implemented in real-world clinical practice populations. For example, these databases were used to develop COVID severity models that guided interventions such as who to prioritize vaccinating during the pandemic. However, the clinical setting and observational databases often differ in the types of patients (case mix), and it is a nontrivial process to identify patients with medical conditions (phenotyping) in these databases. In this study, we investigate how sensitive a model's performance is to the choice of development database, population, and outcome phenotype.

Methods We developed > 450 different logistic regression models for nine prediction tasks across seven databases with a range of suitable population and outcome phenotypes. Performance stability within tasks was calculated by applying each model to data created by permuting the database, population, or outcome phenotype. We investigate performance (AUROC, scaled Brier, and calibration-in-the-large) stability and individual risk estimate stability.

Results In general, changing the outcome definitions or population phenotype made little impact on the model validation discrimination. However, validation discrimination was unstable when the database changed. Calibration and Brier performance were unstable when the population, outcome definition, or database changed. This may be problematic if a model developed using observational data is implemented in a real-world setting.

Conclusions These results highlight the importance of validating a model developed using observational data in the clinical setting prior to using it for decision-making. Calibration and Brier score should be evaluated to prevent miscalibrated risk estimates being used to aid clinical decisions.

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

Upcoming Workgroup Calls

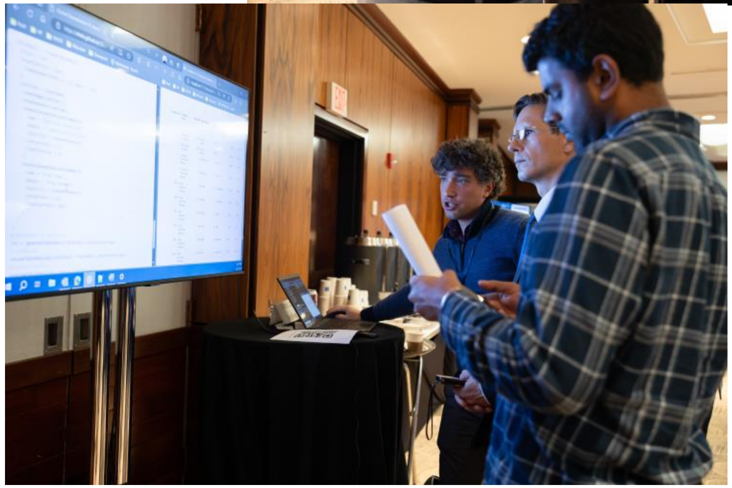
Date	Time (ET)	Meeting
Tuesday	12 pm	ATLAS
Wednesday	9 am	Oncology Outreach/Research Subgroup
Wednesday	12 pm	Latin America
Thursday	9:30 am	Network Data Quality
Thursday	7 pm	Dentistry
Friday	9 am	Phenotype Development and Evaluation
Monday	9 am	Vaccine Vocabulary
Monday	10 am	Africa Chapter
Monday	10 am	The Getting Started Subgroup
Monday	11 am	Book of OHDSI
Tuesday	9:30 am	CDM Survey Subgroup

DevCon 2025: April 25

Agenda

9:00 – 9:15am ET · Welcome & Introduction
- Paul Nagy, Johns Hopkins University

9:15 – 11:30am ET · OHDSI Projects Lightning Talks
- Stabilizing Gaia Core – Robert Miller, Miller Data Solutions
- CustomVocabularyBuilder – Jared Houghtaling, Tufts University
- CohortConstructor – Núria Mercadé-Besora, University of Oxford
- Updates on Strategus – Anthony Sena, Johnson & Johnson
- Experiences with SQLMesh/CICD integration with Databricks – Vishnu Chandrabalan, Lancashire Teaching Hospitals NHS Foundation Trust
- Updates from the Technical Advisory Board – Frank Defalco, Johnson & Johnson



11:30 – 12:30pm ET · Developer dialogue: Dev ops, DBT and, of course, LLMs
Moderator: Katy Sadowski, Boehringer Ingelheim
- Eduard Korchmar, EPAM Systems
- Egill Fridgeirsson, Erasmus MC
- Martin Lavallee, Boehringer Ingelheim
- Lawrence Adams, Artificial Intelligence Centre for Value Based Healthcare

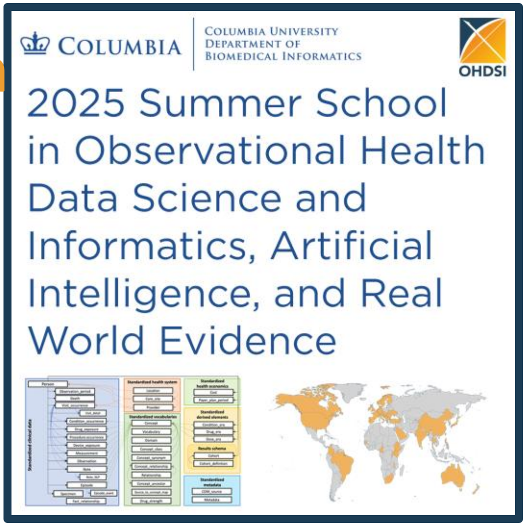
12:30 – 1:00pm ET · Break

1:00 – 2:00pm ET · Sustainable Open-Source Ecosystems Panel
Moderator: Paul Nagy, Sean O'Reilly
- Data4Life – Peter Hoffmann
- The Hyve – Jan Blom/Wouter Franke
- Cognome – James Green

ohdsi.org/DevCon2025

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

Save The Date!

The submission deadline for the 2025 Global Symposium Collaborator Showcase is July 1.

The showcase will be accepting both posters and software demos, as well as interest in hosting lightning talks. More information on the symposium, including abstract submission and registration links, will be available soon.

#OHDSISocialShowcase This Week

Monday

PHederation - the federated network of Pulmonary Hypertension registries

(Eva-Maria Didden, Valerie van Baalen, Michel van Speybroeck, Monika Brand)

PHederation

The Federated Network of Pulmonary Hypertension Registries

PRESENTER: **Michael Briganti**

CO-AUTHORS: Eva-Maria Didden, Valerie van Baalen, Michel van Speybroeck, Monika Brand

INTRO

- Pulmonary Arterial Hypertension (PAH) is a rare subgroup of Pulmonary Hypertension (PH).
- Real-world evidence (RWE) generation in rare diseases is often restricted due to small patient numbers, geographic distribution, and limited data access.
- Federated Data Networks (FDNs) can bring together multiple fit-for-purpose data sources.
- PHederation is a public-private partnership connecting disease-specific clinical data sources for enhanced, transparent, and reproducible research in PH (Ref.1).

METHODS

PHederation data network

- Each database custodian is in control of patient-level data and responsible for privacy, ethics and legal compliance.
- The IT infrastructure consists of a central hub and local software instances, for controlled exchange of queries and result (Fig.1).

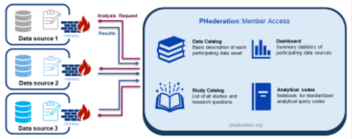


Figure 1. PHederation Portal and mechanisms for collaboration.

Data harmonization

All data were originally mapped to OMOP CDM v.5.3.1 (Ref. 2-3) and undergo refreshes to OMOP CDM v.5.4.1.

Study conduct

1. Select databases from Data Catalogue (Table 1) and perform fit for purpose evaluations.
2. Create study team to develop protocol and analysis plan.
3. Translate analysis plan into executable queries.
4. Distribute queries, execute analysis, share aggregate results, and potentially conduct meta-analysis.
5. Interpret results, write study report, publish.

PHederation

sets a blueprint

for future disease-specific

federated data networks.

Visit us at PHederation.org!

RESULTS

Title of the database	Description	Observation period	Number of PH and PAH patients	Regions	Source data format/ collection
Canadian PH Registry (CPHR)	Prospective PH patient registry	2017 - ongoing	PH – 1´995 PAH – 1´076	Canada	PAHTool
EXPOSURE (EUPAS19085)	Registry of PAH patients newly treated with either Uptravi or another PAH-specific therapy	2017 - ongoing	PAH – 2'354	Europe, Canada	CDISC SDTM
OPUS (NCT02126943)	Opsumit drug registry	2014 - 2018	PH – 2'674 PAH – 2'208	USA	CDISC SDTM
ORPHEUS (NCT03197688)	Opsumit user medical chart review to supplement OPUS	2013 - 2017	PH – 3'031 PAH – 2'410	USA	CDISC SDTM
Porto center of Portuguese PH network	Northern Region Portuguese PH registry	2001 - ongoing	PH – 578 PAH – 216	Portugal - Northern Region	PAHTool
SPHERE (NCT03278002)	Selexipag drug registry	2016 - 2020	PH – 829 PAH – 759	USA	Registry-specific
Stanford clinical PH database	PH Registry	2004 - ongoing	PH – 1´189 PAH – 987	USA – Western Region	Registry-specific

Table 1. PHederation Data Catalogue.

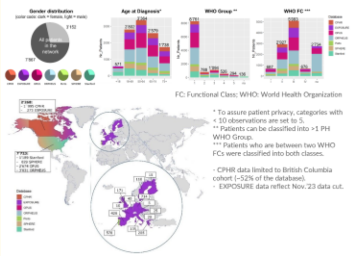


Figure 2. PHederation Dashboard: Patient demographics and disease characteristics at enrolment, and geographic coverage.

CONCLUSION & OUTLOOK

PHederation established a network of databases

- of diverse purpose and origin
- with the goal of advancing scientific knowledge in PH through distributed data sources and analytics, harmonization, and automation.

First PHederation network study (ongoing):

- Objective: Drug utilization assessment of ERAs and PDE-5 inhibitors in newly-diagnosed PAH patients.
- Goal: To complement DARWIN-EU's EUPAS106052 (Ref.4) with evidence from a disease-specific FDN.

REFERENCES:

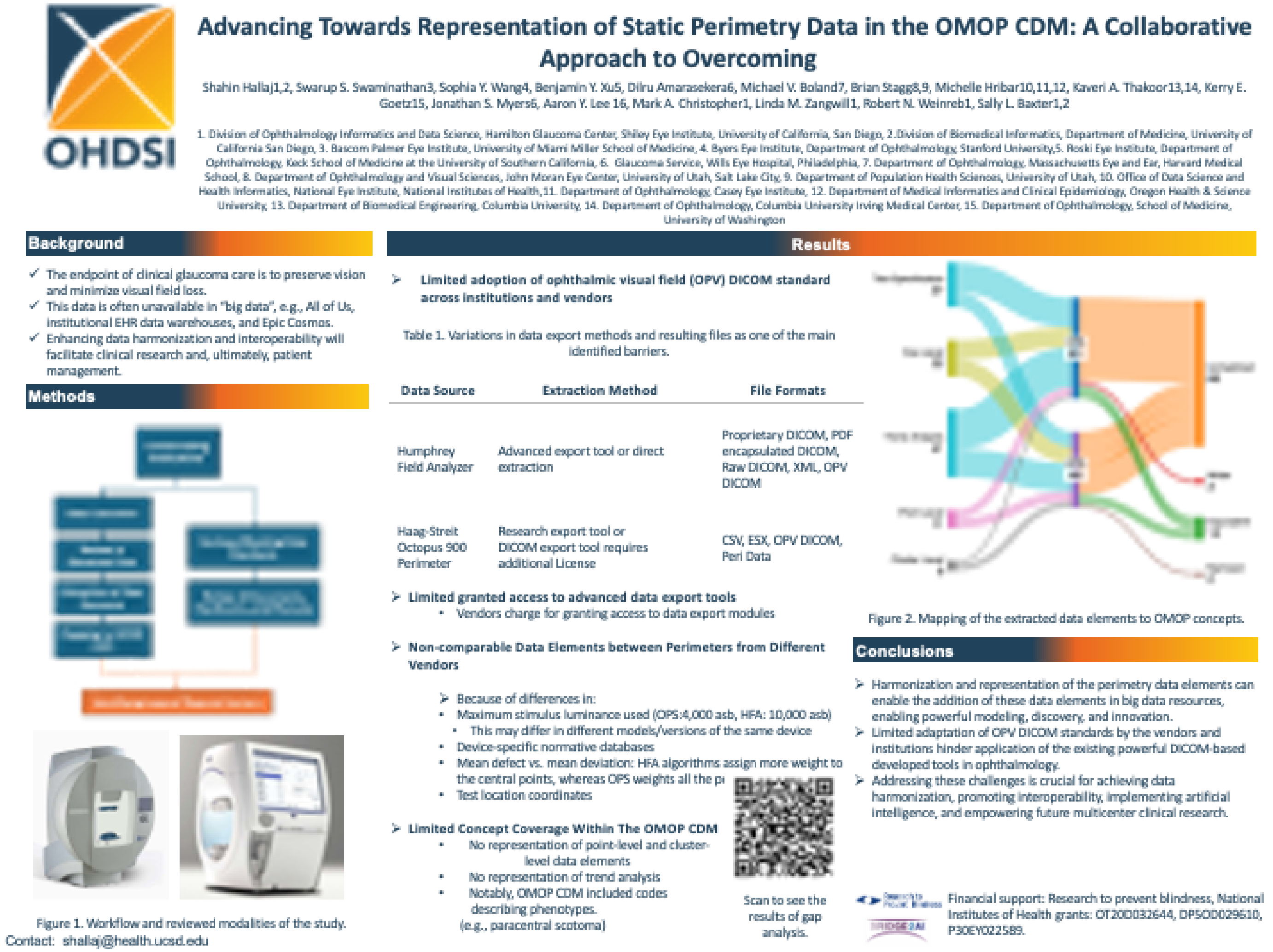
1. Transparency and reproducibility through disease-specific FDNs: https://pubmed.ncbi.nlm.nih.gov/38556812/
2. Standardizing PH registry data to the OMOP Common Data Model: https://pubmed.ncbi.nlm.nih.gov/34727871/
3. Handbook for PH registries to OMOP CDM conversion: https://github.com/OHDSI/ETL--PulmonaryHypertensionRegistries
4. EUPAS106052: https://catalogues.ema.europa.eu/node/3797/administrative-details

#OHDSISocialShowcase This Week

Tuesday

Gap Analysis of Static Automated Perimetry Concept Representation in OMOP CDM

(Shahin Hallaj, William Halfpenny, Niloofar Radgoudarzi, Michael V. Boland, Swarup S. Swaminathan, Sophia Y. Wang5, Benjamin Y. Xu, Dilru C. Amarasekera, Brian Stagg, Michelle Hribar, Kaveri A. Thakoor, Kerry E. Goetz, Jonathan S. Myers, Aaron Y. Lee, Mark A. Christopher, Linda M. Zangwill, Robert N. Weinreb, Sally L. Baxter)


OHDSI

Advancing Towards Representation of Static Perimetry Data in the OMOP CDM: A Collaborative Approach to Overcoming

Shahin Hallaj1,2, Swarup S. Swaminathan3, Sophia Y. Wang4, Benjamin Y. Xu5, Dilru Amarasekera6, Michael V. Boland7, Brian Stagg8,9, Michelle Hribar10,11,12, Kaveri A. Thakoor13,14, Kerry E. Goetz15, Jonathan S. Myers6, Aaron Y. Lee 16, Mark A. Christopher1, Linda M. Zangwill1, Robert N. Weinreb1, Sally L. Baxter1,2

1. Division of Ophthalmology Informatics and Data Science, Hamilton Glaucoma Center, Shiley Eye Institute, University of California, San Diego, 2. Division of Biomedical Informatics, Department of Medicine, University of California San Diego, 3. Bascom Palmer Eye Institute, University of Miami Miller School of Medicine, 4. Byers Eye Institute, Department of Ophthalmology, Stanford University, 5. Roski Eye Institute, Department of Ophthalmology, Keck School of Medicine at the University of Southern California, 6. Glaucoma Service, Wills Eye Hospital, Philadelphia, 7. Department of Ophthalmology, Massachusetts Eye and Ear, Harvard Medical School, 8. Department of Ophthalmology and Visual Sciences, John Moran Eye Center, University of Utah, Salt Lake City, 9. Department of Population Health Sciences, University of Utah, 10. Office of Data Science and Health Informatics, National Eye Institute, National Institutes of Health, 11. Department of Ophthalmology, Casey Eye Institute, 12. Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, 13. Department of Biomedical Engineering, Columbia University, 14. Department of Ophthalmology, Columbia University Irving Medical Center, 15. Department of Ophthalmology, School of Medicine, University of Washington

Background

- The endpoint of clinical glaucoma care is to preserve vision and minimize visual field loss.
- This data is often unavailable in "big data", e.g., All of Us, institutional EHR data warehouses, and Epic Cosmos.
- Enhancing data harmonization and interoperability will facilitate clinical research and, ultimately, patient management.

Methods

Figure 1. Workflow and reviewed modalities of the study.

Contact: shallaj@health.ucsd.edu

Results

- Limited adoption of ophthalmic visual field (OPV) DICOM standard across institutions and vendors

Table 1. Variations in data export methods and resulting files as one of the main identified barriers.

Data Source	Extraction Method	File Formats
Humphrey Field Analyzer	Advanced export tool or direct extraction	Proprietary DICOM, PDF encapsulated DICOM, Raw DICOM, XML, OPV DICOM
Haag-Streit Octopus 900 Perimeter	Research export tool or DICOM export tool requires additional License	CSV, ESX, OPV DICOM, Peri Data

- Limited granted access to advanced data export tools
  - Vendors charge for granting access to data export modules
- Non-comparable Data Elements between Perimeters from Different Vendors
  - Because of differences in:
  - Maximum stimulus luminance used (OPS:4,000 asb, HFA: 10,000 asb)
    - This may differ in different models/versions of the same device
  - Device-specific normative databases
  - Mean defect vs. mean deviation: HFA algorithms assign more weight to the central points, whereas OPS weights all the p
  - Test location coordinates
- Limited Concept Coverage Within The OMOP CDM
  - No representation of point-level and cluster-level data elements
  - No representation of trend analysis
  - Notably, OMOP CDM included codes describing phenotypes (e.g., paracentral scotoma)

Scan to see the results of gap analysis.

Figure 2. Mapping of the extracted data elements to OMOP concepts.

Conclusions

- Harmonization and representation of the perimetry data elements can enable the addition of these data elements in big data resources, enabling powerful modeling, discovery, and innovation.
- Limited adaptation of OPV DICOM standards by the vendors and institutions hinder application of the existing powerful DICOM-based developed tools in ophthalmology.
- Addressing these challenges is crucial for achieving data harmonization, promoting interoperability, implementing artificial intelligence, and empowering future multicenter clinical research.

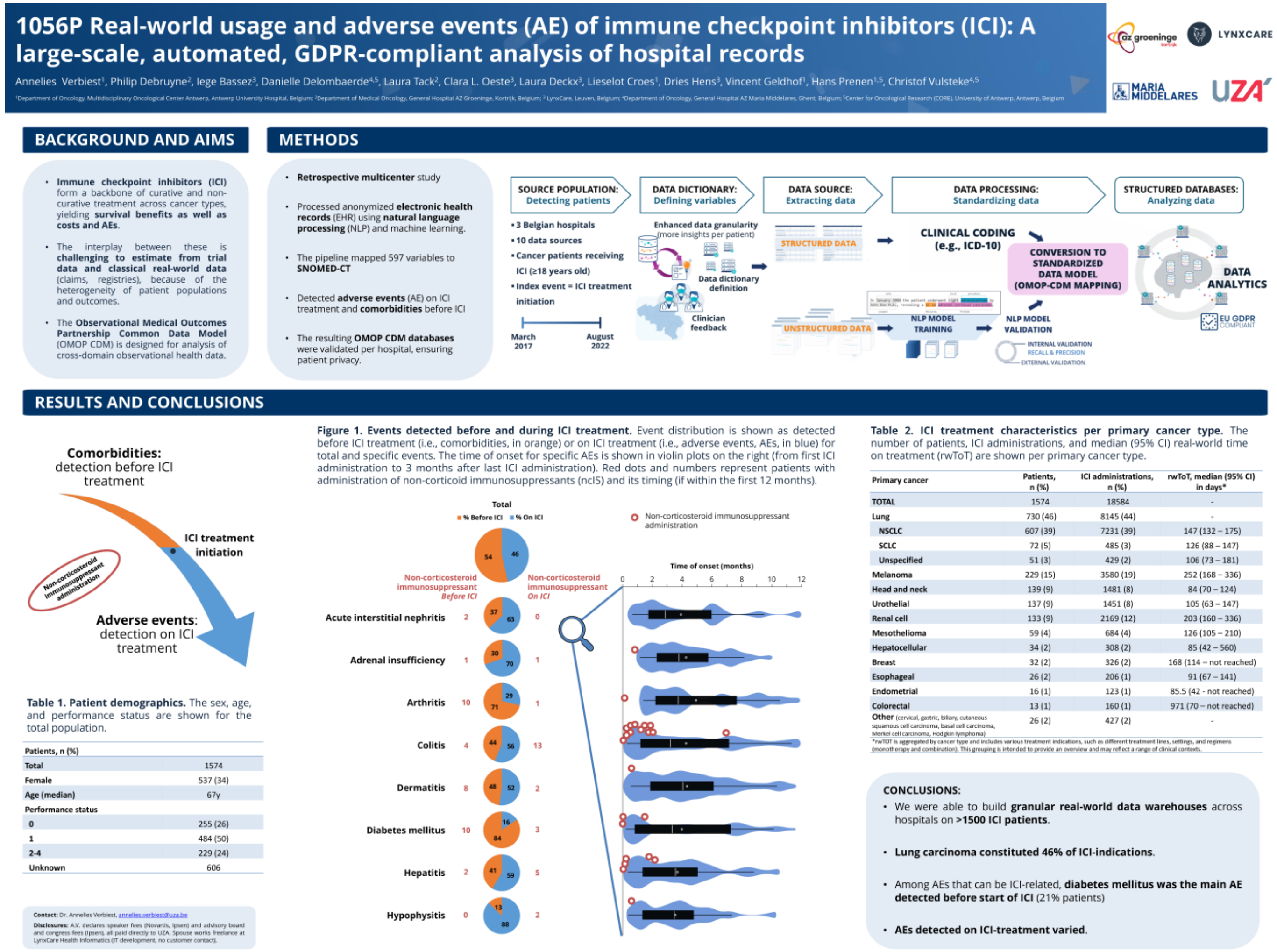
Financial support: Research to prevent blindness, National Institutes of Health grants: OT2OD032644, OP5OD029610, P30EY022589.

#OHDSISocialShowcase This Week

Wednesday

Enhancing Cardiovascular Adverse Event Detection in ICI-Treated Cancer Patients: Lessons Learned from Natural Language Processing Integration with OMOP CDM

(Clara L. Oeste, Danielle Delombaerde, Iege Bassez, Annelies Verbiest, Philip Debruyne, Christof Vulsteke, Dries Hens)



#OHDSISocialShowcase This Week

Thursday

Electrocardiogram-Based Identification of Acute Heart Failure in Chronic Heart Failure: A MIMIC-IV and OMOP-CDM Standardized Approach

(Seung Wook Lee)

***Electrocardiogram-Based Identification of Acute Heart Failure in Chronic Heart Failure:** A MIMIC-IV and OMOP-CDM Standardized Approach*

PRESENTER: **Seung Wook** Lee

INTRO:

- **Acute heart failure (AHF)** is a rapid worsening of heart failure due to fluid overload, leading to symptoms like shortness of breath and swelling. This is especially concerning for chronic heart failure (CHF) patients, who face **higher risks of hospitalization and mortality**.
- Diagnosing AHF typically involves imaging and lab tests, but these require medical attention or specialized equipment.
- This study explores using **ECG data** as a non-invasive tool **to identify AHF in CHF patients**, leveraging data from the MIMIC-IV database standardized to OMOP-CDM guidelines.

METHODS

1. Dataset: MIMIC-IV and MIMIC-IV-ECG from PhysioNet, standardized to OMOP-CDM. (Observation period: 2008-2019)
2. Inclusion Criteria: those with diagnosis of CHF, those with ECG measured within 12 hours of ED admission, those with concurrent diagnosis of AHF during observation period
3. Study Population: Final cohort included 6,988 CHF patients with 40,239 ECGs
4. Baseline ECG and demographic features were idenitifed. ECG differences were computed for 33,251 studies after excluding 6,988 baseline ECGs.
5. Final Analysis Group: 21,315 non-AHF ECGs. 11,936 AHF ECGs.
6. Analysis: XGBoost model developed using age, gender, and ECG-derived features. Performance measured using AUROC and AUPRC.

ECG data may assist in identifying **acute heart failure** in **chronic heart failure** patients, offering a potential tool to support diagnosis.

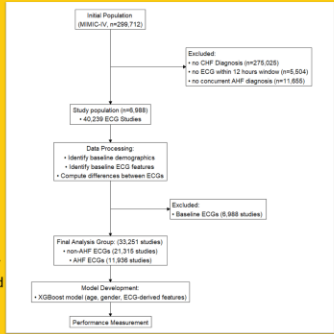


Figure 1: Consort Diagram for study population and inclusion/exclusion criteria

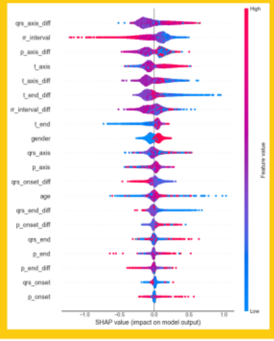


Figure 2: Feature importance for XGBoost Model in SHAP Values, ordered based on impact on model output

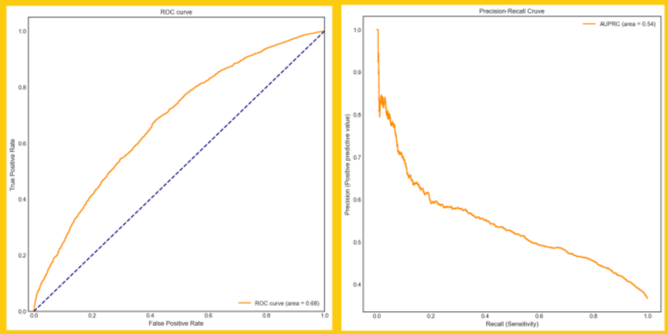


Figure 3: Receiver Operating Characteristic Curve (ROC Curve, left), and Precision-Recall Curve (PRC, right) for XGBoost model built based on demographics and ECG-derived features to identify AHF in CHF population

Results

- Analyzed 6,988 chronic heart failure (CHF) patients, yielding 11,936 ECGs associated with AHF events and 21,315 without AHF.
- Using an XGBoost model, we achieved an AUROC of 0.68 and an AUPRC of 0.54, with a recall of 0.80 and precision of 0.45.
- Key predictors of AHF included QRS axis, RR interval, P wave axis, and T wave axis, reflecting underlying cardiac stress and electrical disturbances commonly seen in AHF.

Conclusion

- ECG data combined with machine learning may assist in identifying acute heart failure (AHF) in chronic heart failure (CHF) patients.
- The model's moderate AUROC and AUPRC suggest it should be used alongside other diagnostic tools.
- High recall indicates its potential for identifying patients at risk of AHF.
- Further validation using external datasets is needed to confirm robustness and generalizability.

Limitations

- The model was trained on a single cohort (MIMIC-IV database), limiting generalizability.
- External validation in diverse populations is necessary.
- The study used only ECG-derived features and demographic data, excluding other clinical information that may improve accuracy.
- Moderate AUROC (0.68) and low precision indicate this method should supplement, not replace, other diagnostic approaches in clinical practice.

LinkedIn Profile

Abstract (pdf)

Seung Wook Lee,
Department of Medicine,
MetroWest Medical Center,
Framingham, MA, USA
seungwooklee94@gmail.com

MetroWest Medical Center

OHDSI

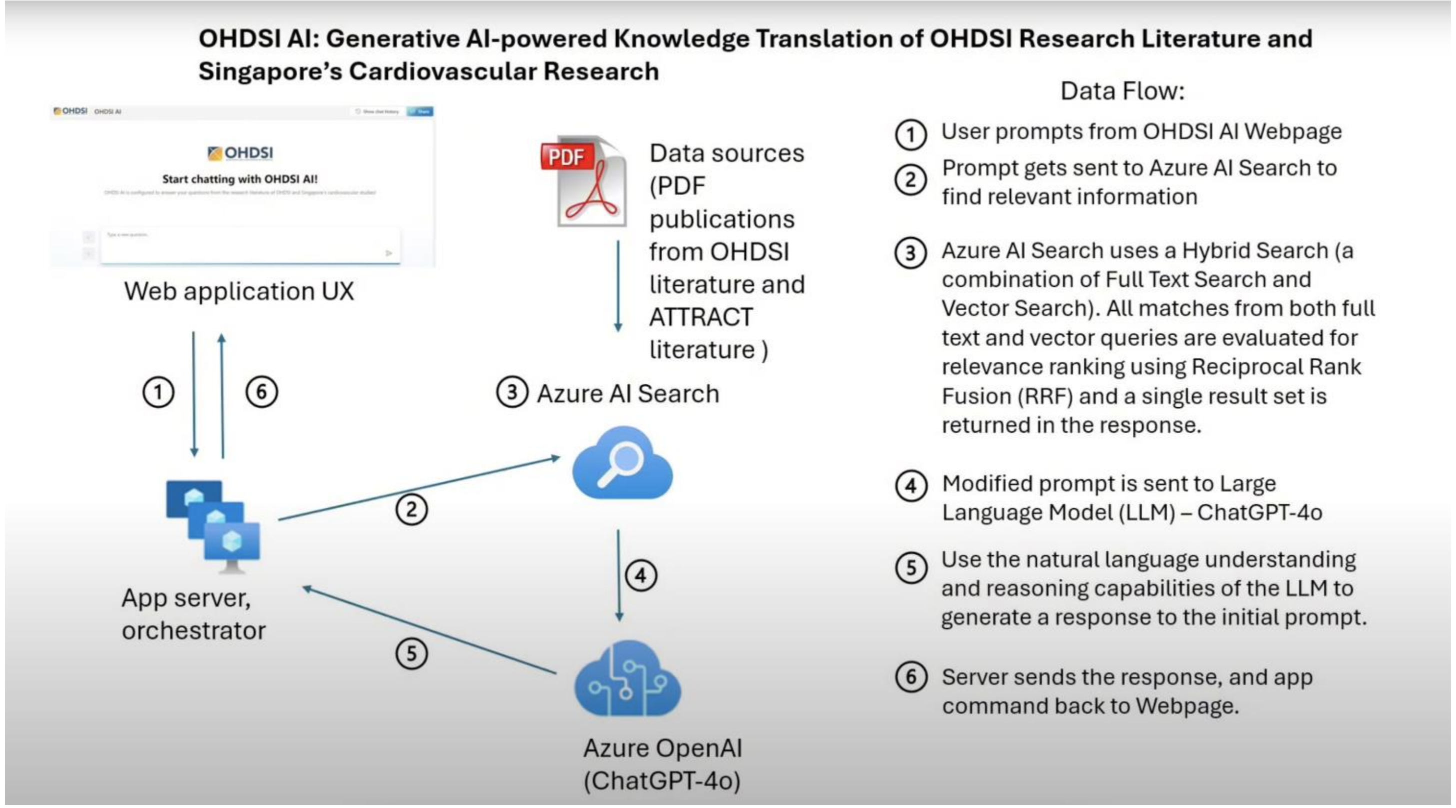
@OHDSI www.ohdsi.org #JoinTheJourney ohdsi

#OHDSISocialShowcase This Week

Friday

OHDSI AI: Generative AI-powered Knowledge Translation of OHDSI Research Literature and Singapore's Cardiovascular Research

(Maisie Ng, Cindy Ho, Li Ting Ang, Hang Png, Shuen Lin Tan, Estella Ye, Ismail Mohd, Mengling Feng, Sebastian Maurer-Stroh, Johan G Eriksson, Mukkesh Kumar)



Where Are We Going?

Any other announcements of upcoming work, events, deadlines, etc?

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

April 22: Estimation and Prediction

George Hripcsak

Vivian Beaumont Allen Professor of Biomedical Informatics, Columbia University

Marc Suchard

Professor of Biostatistics, Biomathematics, & Human Genetics, UCLA

Ross Williams

Assistant Professor, Erasmus University Medical Centre

Egill Fridgeirsson

Scientific Researcher/Postdoc, Erasmus University Medical Centre

Jenna Reps

Associate Director, Observational Health Data Analytics, Johnson & Johnson

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