Holiday Farewell to 2025

OHDSI Community Call

Dec. 16, 2025 • 11 am ET

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Upcoming Community Calls

Date	Topic
Dec. 16	Holiday Farewell To 2025
Dec. 23	No Meeting
Dec. 30	No Meeting
Jan. 6	No Meeting
Jan. 13	Where Can We Go Together in 2026?

Three Stages of The Journey

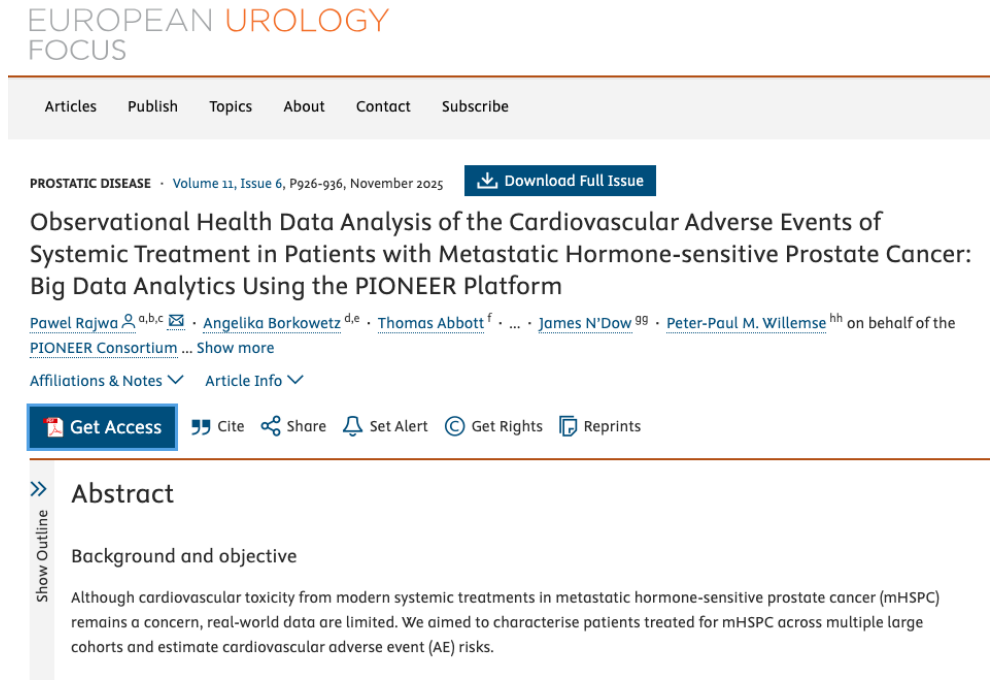
Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Pawel Rajwa, Angelika Borkowetz, Thomas Abbott, Andrea Alberti, Katharina Beyer, Anders Bjartell, James T Brash, Andrew Chilelli, Eleanor Davies, Bertrand De Meulder, Tamas Fazekas, Asieh Golozar, Ayman Hijazy, Andreas Josefsson, Veeru Kasivisvanathan, Raivo Kolde, Daniel Kotik, Michael S Leapman, Marcin Miszczyk, Rossella Nicoletti, Peter Prinsen, Sebastiaan Remmers, Maria J Ribal, Juan Gómez Rivas, Lara Rodriguez-Sanchez, Monique J Roobol, Emma Smith, Robert Snijder, Carl Steinbeisser, Hein V Stroomberg, Giorgio Gandaglia, Philip Cornford, Susan Evans-Axelsson, James N'Dow, Peter-Paul M Willemse; and the PIONEER Consortium** on the publication of **Observational Health Data Analysis of the Cardiovascular Adverse Events of Systemic Treatment in Patients with Metastatic Hormone-sensitive Prostate Cancer: Big Data Analytics Using the PIONEER Platform** in *European Urology Focus*.


EUROPEAN UROLOGY FOCUS

Articles Publish Topics About Contact Subscribe

PROSTATIC DISEASE · Volume 11, Issue 6, P926-936, November 2025 Download Full Issue

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

Pawel Rajwa [a,b,c] · Angelika Borkowetz [d,e] · Thomas Abbott [f] · ... · James N'Dow [gg] · Peter-Paul M. Willemse [hh] on behalf of the PIONEER Consortium ... Show more

Affiliations & Notes Article Info

Get Access Cite Share Set Alert Get Rights Reprints

Show Outline

Abstract

Background and objective

Although cardiovascular toxicity from modern systemic treatments in metastatic hormone-sensitive prostate cancer (mHSPC) remains a concern, real-world data are limited. We aimed to characterise patients treated for mHSPC across multiple large cohorts and estimate cardiovascular adverse event (AE) risks.

OHDSI Shoutouts!

Congratulations to the team of **Ka Hee Yoo, Kyung Joo Lee, Sang Min Lee, Changwoo Han, Rae Woong Park, and Young Tak Jo** on the publication of **Comparative effectiveness of selective serotonin reuptake inhibitors versus serotonin-norepinephrine reuptake inhibitors in the risk of diagnostic conversion from unipolar depression to bipolar disorder** in the *International Journal of Psychiatry in Clinical Practice*.

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/WebAPI
Thursday	8 am	India Community Call
Thursday	12 pm	HADES
Friday	10 am	GIS– Geographic Information System
Friday	10:30 am	Open-Source Community
Monday	9 am	Africa Chapter

ATHENA Survey

Athena user survey

Help us understand how to make Athena better

When you submit this form, it will not automatically collect your details like name and email address unless you provide it yourself.

* Required

1. If you are open to follow-up about your feedback, please provide your email address

Enter your answer

2. How do you use athena.ohdsi.org? *

- [ ] Search concepts
- [ ] Download current version of vocabularies
- [ ] Download previous version of vocabularies
- [ ] Other

3. What do you athena.ohdsi.org for? *

- [ ] ETL data
- [ ] Search concepts to create mappings
- [ ] Search concepts for concept sets (value sets, code lists)
- [ ] Translate concepts to other languages/find translations
- [ ] Use as knowledge graph outside of OMOP CDM
- [ ] Other

4. On average, how often do you access athena.ohdsi.org? *

- ◯ Every day
- ◯ Once a week
- ◯ A few times a month
- ◯ Once a month
- ◯ Once 6 month or less
- ◯ Other

5. If there was an Athena API, how would you use it and what would you use it for? *

Enter your answer

6. Anything else you'd like to tell us? *

Enter your answer

You can print a copy of your answer after you submit

Submit

Columbia DBMI Summer School

The 2026 Summer School in Observational Health Data Science & Informatics, AI, and Real World Evidence

COLUMBIA | COLUMBIA UNIVERSITY DEPARTMENT OF BIOMEDICAL INFORMATICS

OHDSI OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

June 22–26, 2026, Columbia Biomedical Informatics

The Columbia OHDSI Summer School provides health professionals, researchers, and industry practitioners with an immersive, hands-on training to working with real-world health data and generating real-world evidence (RWE). Participants will explore the types of healthcare data captured during routine clinical care—such as electronic health records and administrative claims—and learn how to standardize these data using the OMOP Common Data Model to support collaborative, distributed research as part of a data network.

Over the course of the week, participants will engage with three real-world analytic use cases:

- **Clinical characterization** – using descriptive epidemiology to study disease natural history and treatment patterns
- **Population-level estimation** – applying causal inference to assess drug safety and comparative effectiveness
- **Patient-level prediction** – leveraging machine learning for early disease detection and precision medicine

Participants will be guided through the full RWE study lifecycle: from designing observational studies tailored to each use case, to applying open-source tools form the OHDSI community, and executing analyses across real-world data sources.

The curriculum combines foundational lectures on analytical methods with hands-on, interactive, faculty-led group exercises. In addition, participants will have dedicated time to develop and advance their own study concepts with personalized feedback and mentoring.

Oxford Summer School Registration Opens

Oxford Summer School 2026: Real World Evidence using the OMOP Common Data Model

COURSE DIRECTORS

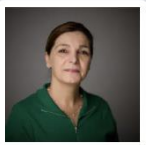
Daniel Prieto-Alhambra

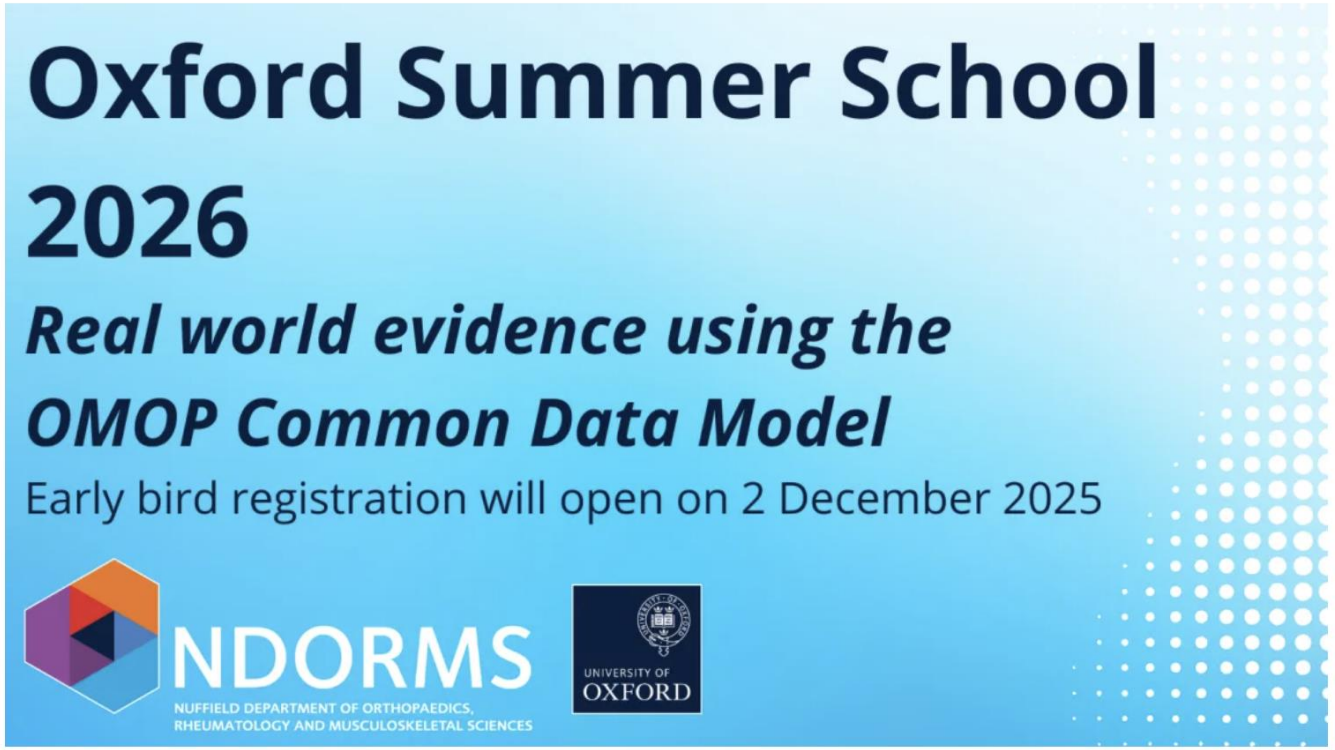
Professor of Pharmaco- and Device Epidemiology

COURSE ADMINISTRATOR

Mahkameh Mafi

Personal Assistant to Professor Prieto-Alhambra





2026 Global Symposium

The 2026 OHDSI Global Symposium will return to the Hyatt Regency Hotel in New Brunswick, N.J., on **Oct. 20-22.**

2026 Global Symposium

2026 OHDSI Global Symposium Call for Plenary Sessions

Symposium plenaries provide opportunities to share innovative, community-developed content to empower researchers to generate reliable real-world evidence. The community is currently seeking proposals for our #OHDSI2026 plenaries. These sessions will be 60 minutes in duration and must touch on at least two of following pillars of our community:

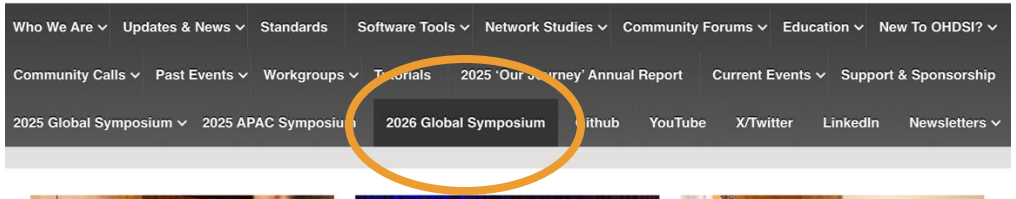
- Open community data standards
- Methodological research
- Open-source development
- Clinical applications

Plenary sessions must also involve three or more on-stage participants across at least two organizations. Sessions may include a combination of keynote talks, panel discussions, interactive activities, and more. We strongly encourage using multiple formats and synthesizing completed research, current perspectives and future calls-to-action to maximize community engagement.

The deadline for proposal submissions is January 30, 2026. Please use the link below to submit your proposal by answering the following questions:

- Name(s) of plenary session organizers:
- Your email address(es):
- Short (2,500 character max) description / abstract of your proposed session:
- Which pillars are you targeting:
- One sentence “pitch” of your session to excite the community:
- Names and roles of individuals who have tentatively agreed to participate in your session:

Deadline to submit proposals for #OHDSI2026 plenaries or tutorials is Jan. 30, 2026!



2026 OHDSI Global Symposium

Oct. 20-22 • New Brunswick, N.J. • Hyatt Regency Hotel

2026 OHDSI Global Symposium Call for Tutorials

Tutorial sessions aim to deliver educational content, led by community members who wish to train our global collaborators on scientific, technical, and other skills that can support advancing OHDSI’s mission and the effective use of real-world data and the generation and dissemination of reliable real-world evidence. Examples of prior tutorials offered are provided here: https://www.ohdsi.org/tutorials.

Tutorial sessions are 4 hours in duration. Registrants for your tutorial will be requested to pay a registration fee. The fees will be used to offset the costs of the symposium and other OHDSI expenses. Sessions may include a combination of talks, interactive activities, and more. We strongly encourage using multiple formats to maximize community engagement. Your session must include at least three people from at least two different organizations.

The deadline for tutorial proposal submissions is January 30, 2026. Please use the link below to submit your proposal by answering the following questions:

- Name(s) of tutorial session organizers:
- Your email address(es):
- Short (2,500 character) description / abstract of your proposed session:
- Names and roles of individuals who have tentatively agreed to participate in your session:

2026 Europe Symposium

The 2026 OHDSI Europe Symposium returns to Rotterdam next year and will be held **April 18-20.**

The deadline for abstract submissions will be Feb. 6, 2026.

#OHDSISocialShowcase This Week

Monday

Preserving the DNA of Clinical Intent: Integrating IMO Health API services into OHDSI Extract-Transform-Load Process

(Evan Sholle, David Haines, Chandan Ravishankar, Tejaswini Viswanath, Merlin Simoes, Daniel Timke, Sajjad Abedian, Frank Naeymi-Rad)

Preserving the DNA of Clinical Intent
Integrating IMO Health API services into OHDSI Extract-Transform-Load Process

PRESENTER: **Evan Sholle**

INTRO:
While the OHDSI community has established a well-validated library of techniques for generating evidence from real-world data, the value of data sets mapped to the model is constrained by data quality issues[1] attributable to a host of factors, including factors derived from source data systems and those introduced by the mapping process.
Specifically, within the Condition domain, mapping physician documented clinical terms ("interface terms") to OHDSI standard terms often involves a "step down" into a less granular intermediate terminology - ICD-10 - prior to mapping to SNOMED.[2]
We hypothesized that by incorporating IMO Health terminology services into the ETL process and mapping IMO-derived interface terminology items directly to SNOMED, rather than stepping through ICD-10, we could preserve diagnostic granularity, better reflect physician intent, and improve the performance of computable phenotyping techniques executed against the resultant data set.

METHODS

1. ETL Comparison:
   As illustrated in Figure 1, we developed two pipelines for mapping EHR diagnoses to the OHDSI CDM: a "legacy" ETL representative of current standard practice reliant on ICD-10-CM → SNOMED mapping, and an "Enhanced" ETL using IMO Precision Normalize™ API for direct interface-term → SNOMED mapping with bidirectional ICD-10-CM crosswalks to derive the *condition_source_concept_id*.
2. Patient Impact Analysis:
   After conducting the ETL, we measured the number of unique interface terminology items present within the source data. We then calculated the number of unique SNOMED and ICD-10 codes these terminology items mapped to for each of the two techniques. We also measured the impact of the different ETL techniques on specific ICD-10 codes that were highly impacted by the change.

Mapping diagnoses into the CDM through ICD-10 incurs real data granularity loss! Using IMO Health's suite of terminology services, we can map clinical terms directly to preserve important diagnostic details.

Take a picture to download the full paper

Results
As detailed in Figure 2, the Enhanced ETL technique resulted in a significantly more granular data set. The 162,501 unique interface terminology items in the source data set were mapped to 45,561 unique SNOMED codes and 29,751 ICD-10 codes using the Enhanced ETL technique, while the legacy ETL technique resulted in only 11,389 unique SNOMED codes and 29,348 ICD-10 codes.

Figure 1

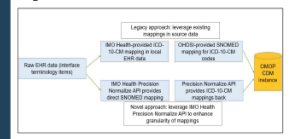


Figure 2

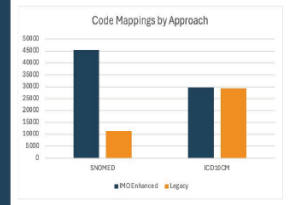


Table 1

ICD-10 code	Diagnosis name	Count of patients with legacy ETL method	Count of patients with enhanced ETL method
Z40.00	Encounter for prophylactic removal of unspecified organ	3	6,330
E65.50	Unspecified deficiencies of globe	1	8510
Z00.71	Encounter for examination for period of delayed growth in childhood with abnormal findings	0	783
H15.9	Unspecified disorder of globe	0	6922
S76.001A	Unspecified injury of muscle, fascia and tendon of right hip, initial encounter	1	162

1. Blackeler C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. J Am Med Inform Assoc. 2021 Sep 18;28(10):2251-2257. doi: 10.1093/jamia/ocab132. PMID: 34313749; PMCID: PMC8449628.
2. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan 1;20(1):144-51. doi: 10.1136/amiajnl-2011-000681. Epub 2012 Jun 25. PMID: 22733976; PMCID: PMC3555312.

Evan Sholle, David Haines, Chandan Ravishankar, Tejaswini Viswanath, Merlin Simoes, Daniel Timke, Sajjad Abedian, Frank Naeymi-Rad

#OHDSISocialShowcase This Week

Tuesday

Prediction of Hyperuricemia and Its Association with Renal failure, Cardiovascular Prognosis in Adults with Type 2 Diabetes mellitus

(Sujin Gan, Dong Yun Lee, Rae Woong Park)

Prediction of Hyperuricemia and Its Association with Renal failure, Cardiovascular Prognosis in Adults with Type 2 Diabetes mellitus

Sujin Gan[1,4,5], Dong Yun Lee[2,3], and Rae Woong Park[1,2,4,5]

[1] Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea; [2] Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Korea
[3] Bongdam forest mental health clinic, Hwaseong, Korea; [4] Center for Biomedical Informatics Research, Ajou University Medical Center, Suwon, Korea; [5] BK21 R&E Initiative for Advanced Precision Medicine, Suwon, Korea

Background

- Accurate prediction of disease progression in T2DM is essential for personalized treatment, yet the prognostic value of hyperuricemia as a potential biomarker remains unclear.

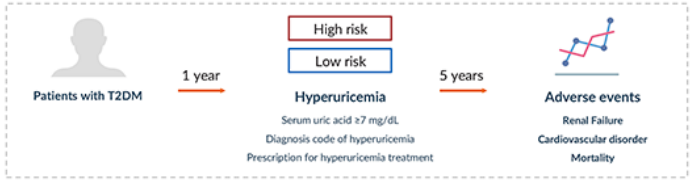
We aimed to:

- Develop and externally validate a machine learning model predicting hyperuricemia within 1 year of T2DM diagnosis.
- Evaluate associations between predicted hyperuricemia and long-term outcomes

Methods

1. Data sources

- OMOP-CDM databases from **4 Korean healthcare institutions** (N = 6,108,232 patients)
  - Model development: Ajou University School of Medicine (AUSOM)
  - External validation: Kangdong Sacred Heart Hospital (KDH), Sejong General Hospital Bucheon (SJBCN), Wonkwang University Hospital (WKUH)



2. Study population

- Target cohort: Adults (≥18 years) with newly diagnosed T2DM
  - Inclusion criteria: ≥ 1 year prior observation; ≥1 serum uric acid measurement within one year post-diagnosis
  - Exclusion criteria: prior hyperuricemia, gout, cancer, renal, cardiovascular, or cerebrovascular disease

3. Model development

- XGBoost, LASSO, Random Forest (in 3-fold Cross validation with 75/25 train-test split)
  - Features: Demographics, Conditions, Drugs, Observations, Procedures (30-day and 365-day windows)
  - Performance metrics: AUROC, AUPRC, F1-score, Accuracy

4. External validation & Long-term Outcomes

- Kaplan–Meier survival estimates and Cox proportional hazards regression
- **Meta-analysis** performed to generate pooled hazard ratios across databases

Conclusions

- This study developed and externally validated a robust prediction model for one-year hyperuricemia in newly diagnosed T2DM patients.
- Importantly, predicted hyperuricemia was associated with a significantly elevated risk of renal failure, cardiovascular complications, and death over 5 years, suggesting that early identification of hyperuricemia risk may inform risk stratification and long-term management strategies in T2DM.
- **Findings highlight hyperuricemia as a prognostic biomarker for adverse outcomes in diabetes.**

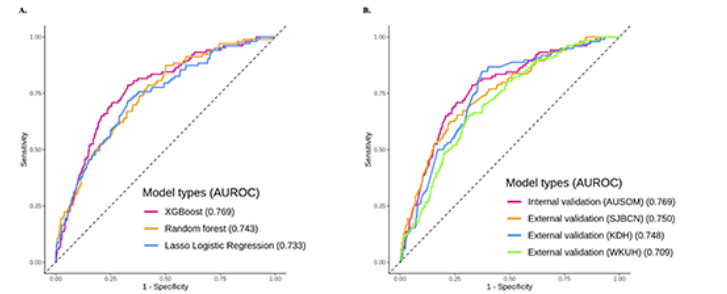
Results

1. Study population characteristics

- In the AUSOM cohort (n = 5,263), 413 patients (7.85%) developed hyperuricemia within one year of T2DM diagnosis. These patients were more likely to be male (67.3%) compared to those without hyperuricemia (47.4%, p < 0.001), but no significant age difference was observed.

2. Model performance

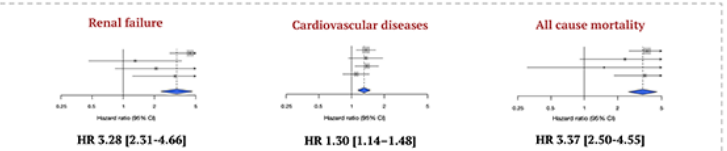
- The XGBoost model achieved the best performance, with an AUROC of 0.769 (95% CI: 0.723–0.815), AUPRC of 0.205, F1-score of 0.278, and accuracy of 0.681. In external validation, robust performance was observed.



- Key predictive features included male sex, use of sulfonylureas or insulin, bacterial infections, hemoglobin levels, and prescriptions of diuretics or anti-tuberculosis drugs.
- Consistent top predictors were identified across XGBoost, LASSO, and Random Forest models.

3. Long term outcomes

- Over a five-year follow-up period, predicted hyperuricemia was strongly associated with adverse outcomes.
- Meta-analysis across four databases confirmed these associations, yielding pooled HRs of 3.28 (95% CI: 2.31–4.65) for renal failure, 1.30 (95% CI: 1.14–1.48) for cardiovascular disease, and 3.37 (95% CI: 2.50–4.56) for mortality.



Fundings: This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MOE)(NO : 2120240615426, BK21 R&E Initiative for Advanced Precision Medicine), 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: RS-2024-00335936), a grant (25212MFDS002) from Ministry of Food and Drug Safety in 2025, and the NRF grant funded by the Korea government(MSIT)(grant numbers RS-2025-20552981 and RS-2025-16070957).

#OHDSISocialShowcase This Week

Wednesday

Real-World Treatment Pathways of Lung Cancer Patients in Taiwan: A Common Data Model Analysis Using TMUCRD

(Nguyen Thi Kim Hien, Thanh-Phuc Phan, Nam Hoai Vo, Muhammad Solihuddin Muhtar, Christianus Heru Setiawan, Septi Melisa, Jason C. Hsu)

Real-World Treatment Pathways of Lung Cancer Patients in Taiwan: A Common Data Model Analysis Using TMUCRD

PRESENTER: Nguyen Thi Kim Hien

INTRO

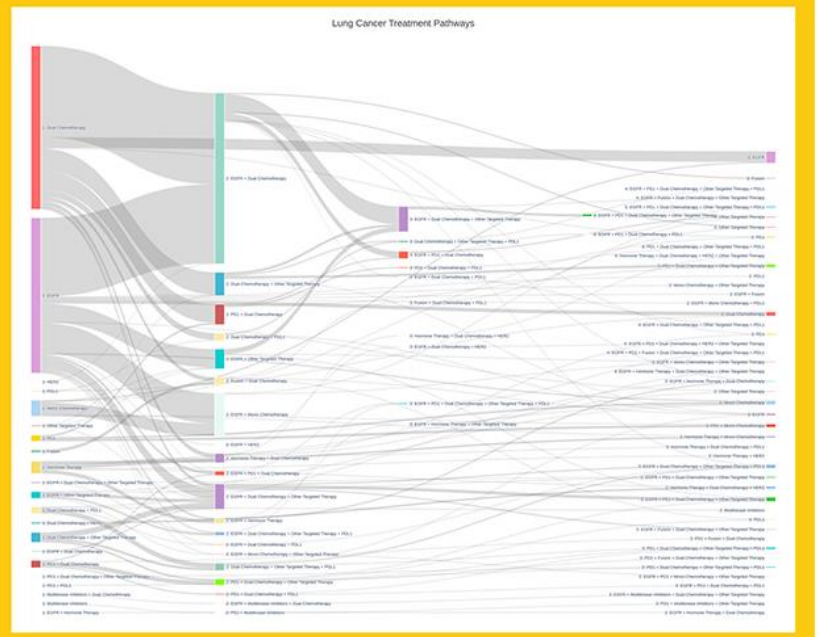
- Lung cancer is the leading cause of cancer deaths globally (1.8M deaths, 18.7% in 2022).
- Clinical trials often don't reflect real-world care.
- Treatment pathway analysis bridges the gap between evidence-based guidelines and actual practice.

METHODS

1. Data source: TMUCRD (Taipei Medical University Hospital, Shuangho, Wanfang) 2005–2020.
2. Format: Transformed into OMOP CDM for standardized OHDSI analysis.
3. Patients: 1,895 adults with primary lung cancer, ≥18 years, treated pharmacologically.
4. Analysis tools: ATLAS, R (TreatmentPatterns), Python.
5. Therapy categories:
   - Mono- & dual chemotherapy
   - EGFR inhibitors
   - Fusion / HER2 / other targeted agents
   - Immune checkpoint inhibitors (PD-1/PD-L1/CTLA-4)
   - Hormonal therapy
6. Pathways: Up to 5 sequential treatment steps visualized.

Chemotherapy and **EGFR inhibitors** dominate **first-line therapy** for lung cancer treatment, but real-world treatment sequences are **highly heterogeneous.**

Substantial patient drop-off beyond second line.



RESULTS

First-line therapies:

- Dual chemotherapy: 39.2%
- EGFR inhibitors: 33.8%
- Mono chemotherapy: 10%

Transitions:

- Patients on dual chemo → often switched to EGFR regimens.
- Patients on EGFR inhibitors → often cycled back to dual chemo.

CONCLUSION

This study highlights the heterogeneous nature of real-world lung cancer treatment pathways, with chemotherapy and EGFR inhibitors as predominant first-line therapies and a gradual integration of targeted and immunotherapies in later lines.

Nguyen Thi Kim Hien, Thanh-Phuc Phan, Nam Hoai Vo, Muhammad Solihuddin Muhtar, Christianus Heru Setiawan, Septi Melisa, Jason C. Hsu*

#OHDSISocialShowcase This Week

Thursday

Empowering Trial by Multi-source Multi-site Real-world Data: A Negative Control-Calibrated Digital Twin Approach

(Dazheng Zhang, Huiyuan Wang, Yiwen Lu, Yong Chen)

OHDSI

Empowering Trial by Multi-source Multi-site Real-world Data: A Negative Control-Calibrated Digital Twin Approach

Dazheng Zhang[a], Huiyuan Wang[a], Yiwen Lu[a], Jingyue Huang[a], Yong Chen[a]

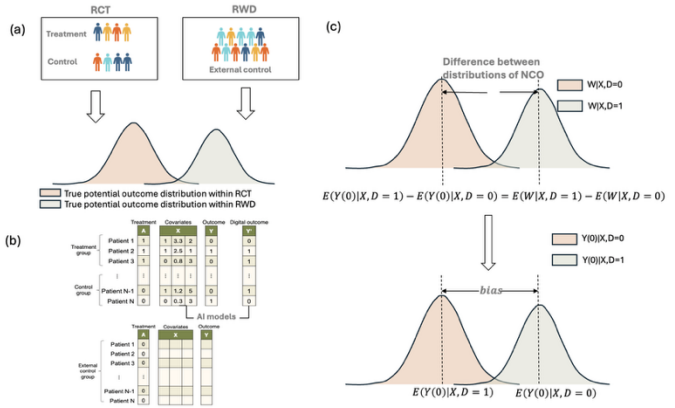
[a] The Center for Health Analytics and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA

Background

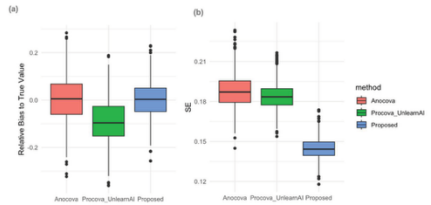
- **Randomized controlled trials** (RCTs) ensure robust causal inference through randomization but often lack generalizability due to strict inclusion criteria, high costs, and limited scalability.
- **Real-world data (RWD)** offer broader population coverage and longitudinal follow-up at lower cost, but are prone to bias due to non-randomization and incomplete data access.
- **Integrating external controls from RWD** can enhance the statistical efficiency of RCTs by reducing required sample sizes and costs.
- **Digital twin technology**, originating from engineering, uses machine learning on high-dimensional RWD to simulate patient-specific counterfactuals and treatment responses. The PROCOVA[1] method applies this by creating virtual control groups to improve trial efficiency[7]. However, discrepancies between trial and real-world populations can introduce bias, necessitating careful calibration.
- We propose a novel framework that corrects **model shift bias** in RCT-RWD integration using digital twins and negative control outcomes.

Method

- We introduce the concept of **NCOs**, outcomes that are theoretically unaffected by treatment assignment, allowing for an empirical estimate of the model shift bias. The observed discrepancy in NCO distributions between RCT and RWD serves as a diagnostic tool for identifying systematic shifts in outcome distributions.
- We utilize predictive modeling within the RWD to construct a **digital twin**—an individualized outcome model for RCT participants. More details about the proposed method can be found the Methods Section. By systematically calibrating for model shift, our method facilitates more robust and generalizable treatment effect estimation in hybrid trial designs.

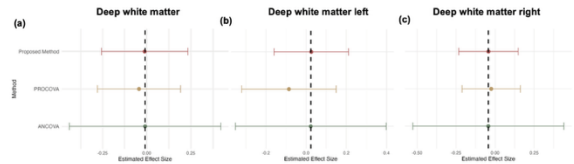


Contact: dazheng.zhang@pennmedicine.upenn.edu and ychen123@upenn.edu



- We illustrate the performance of the proposed method in comparison to baseline methods, including PROCOVA—an approach that does not account for model-shift bias—and ANCOVA.
- The ANCOVA method relies solely on RCT data, making its point estimates the gold standard due to the inherent randomization in RCTs.
- The proposed method achieves a 39.9% relative efficiency gain compared to ANCOVA. In contrast, the PROCOVA method, which is affected by model-shift bias, shows minimal efficiency improvement over ANCOVA.

Results



- As ANCOVA relies solely on RCT data, its estimates serve as the gold standard due to the benefits of randomization.
- The PROCOVA method incorporates external real-world data (RWD) but does not account for model-shift bias, leading to potential inaccuracies.
- The proposed method, by addressing model-shift bias, produces estimates that align closely with ANCOVA while achieving notable efficiency gains. Across all panels, the proposed method consistently yields more precise estimates and avoids the potential distortions observed in PROCOVA.

Conclusion

- In summary, our proposed digital twin framework with negative control outcome calibration effectively addresses model shift bias in RCT-RWD integration.
- Applied to brain imaging data from SPRINT-MIND and iSTAGING, the method achieved treatment effect estimates that closely aligned with RCT-based ANCOVA while offering improved statistical precision.
- This approach enhances the validity and efficiency of hybrid trial designs, paving the way for more reliable real-world evidence for OHDSI community.

Reference

Schuler A, Walsh D, Hall D, Walsh J, Fisher C, Initiative ADN. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *Int J Biostat* 2022; **18**: 329–56

#OHDSISocialShowcase This Week

Friday

How do changes in vocabulary mapping and database release versions affect cohort composition in real-world data?

(Jill Hardin, Evanette Burrows, Azza Shaoibi, Clair Blacketer)

Title: How do changes in vocabulary mapping and database release versions affect cohort composition in real-world data?

PRESENTER: **Jill Hardin**

INTRODUCTION:

1. Challenges exist in assessing effects of vocabulary and database version changes in the analysis observational data.
2. Gaining an understanding of the influence on cohort composition is critical for epidemiological research.
3. This study compares incidence rates (IR) of condition (n=239) and drug (n=198) phenotypes across vocabulary versions and database releases for 13 different data sources.
4. Absolute percent difference was calculated to quantify IR changes.

METHODS:

1. 13 observational health data sources standardized to the OMOP Common Data Model (CDM) (1) v5.4 to derive IRs.
2. Two versions of each data source were created by mapping the native data to standard concepts using two vocabulary versions: February 29, 2024, and February 27, 2025.
3. 437 phenotypes were evaluated: 239 outcomes and 198 drug exposures across the 13 data sources x 2 vocabulary versions.
4. The population at risk (target cohort) was subjects with an observation period between January 1, 2016, and December 31, 2022.
5. ATLAS (2) and the Strategus incidence rate module was used for phenotype development and IR calculation.
6. OHDSI best practices was used to develop the phenotypes.
7. The absolute percent difference between IRs for each vocabulary and database release version was calculated using the formula:

$$\frac{(IR\_new\_vocab - IR\_prior\_vocab)}{IR\_prior\_vocab} \times 100$$

8. Per 100 person years unit used

RESULTS:

1. Most phenotypes (31-85%) showed absolute percent differences of 0-5%.
2. Larger changes (>5%) were seen in Optum EHR (12.6%) and CPRD (7.3%).
3. Phenotypes with greatest differences: heavy menstrual bleeding in Optum EHR & CPRD, and epoprostenol in JMDC.

Table 1: The number (%) of phenotypes by categories of absolute percent difference between incidence rates.

Database	The number (%) of phenotypes by category of absolute percent difference in incidence rates											
	>0-5		>5-10		>10-20		>20-50		>50+		NA*	
	N	%	N	%	N	%	N	%	N	%	N	%
CPRD	261	59.7%	13	3.0%	12	2.7%	6	1.4%	1	0.2%	144	32.9%
France Disease Analyzer	205	46.9%	0	0.0%	2	0.5%	5	1.1%	2	0.5%	223	51.0%
German Disease Analyzer	278	63.6%	5	1.1%	2	0.5%	5	1.1%	3	0.7%	144	32.9%
Health Verity CC	370	84.7%	10	2.3%	5	1.1%	3	0.7%	0	0.0%	49	11.2%
JMDC	277	63.4%	6	1.4%	7	1.6%	6	1.4%	7	1.6%	134	30.6%
LPD Australia	136	31.1%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	301	68.8%
CCAE	367	84.0%	2	0.5%	11	2.5%	2	0.5%	0	0.0%	55	12.6%
MDCR	351	80.3%	3	0.7%	8	1.8%	1	0.2%	0	0.0%	74	16.9%
MDCD	369	84.4%	3	0.7%	7	1.6%	3	0.7%	0	0.0%	55	12.6%
Optum EHR	342	78.3%	24	5.5%	22	5.0%	7	1.6%	2	0.5%	40	9.1%
Optum DOD	372	85.1%	2	0.5%	8	1.8%	3	0.7%	0	0.0%	52	11.9%
Optum SES	372	85.1%	2	0.5%	8	1.8%	3	0.7%	0	0.0%	52	11.9%
Premier	283	64.8%	8	1.8%	6	1.4%	1	0.2%	0	0.0%	139	31.8%

*NOTE: NA bin is created when the incidence rate on new and old vocabulary are equal to 0

Phenotype composition is stable across data releases and vocabulary updates.

Figure 1. Condition phenotypes with absolute percent difference values >5% by database

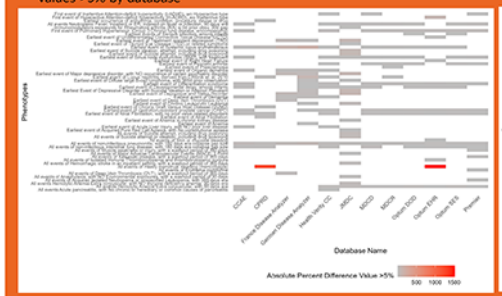
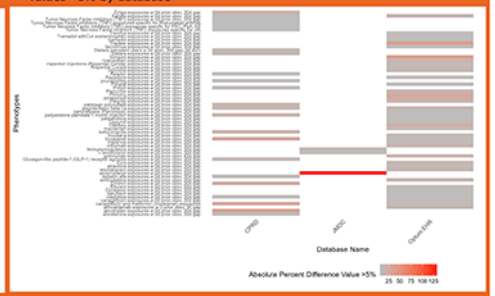


Figure 2. Drug exposure phenotypes with absolute percent difference values >5% by database



Take a picture to download the poster

RESULTS cont:

1. Optum EHR & CPRD use SNOMED; others use ICD10.
2. JMDC data showed drug mappings added across releases, especially for epoprostenol, mapped to English codes.

CONCLUSIONS:

1. Phenotype composition is stable across data releases and vocabulary updates.
2. Certain phenotypes may be sensitive to vocabulary or release changes due to shifts in SNOMED concepts or drug mappings.
3. Data sources using native SNOMED are more affected by hierarchy changes.
4. This study highlights the need to monitor and understand the impacts of vocabulary and data releases in real-world data analysis.
5. The approach provided can help develop standardized methods to assess these effects on phenotype definitions.

FUTURE WORK:

1. Incorporate this approach into a standardized tool

REFERENCES:

1. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015;22(3):553-64.
2. https://github.com/OHDSI/Atlas
3. Sena A, Schuemie M, Gilbert J (2025). Strategus: Coordinate and Execute OHDSI HADES Modules. R package version 1.3.0, https://github.com/OHDSI/Strategus, https://ohdsi.github.io/Strategus.

Jill Hardin[1,2], Evanette Burrows[1,2], Azza Shaoibi[1,2], Clair Blacketer[1,2]
[1]Observational Health Data Analytics, Global Epidemiology, Johnson & Johnson Research and Development, Raritan, NJ, USA; [2]Observational Health Data Sciences and Informatics, New York, NY, USA;

Johnson&Johnson Global Epidemiology Organization

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?

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