# The Book of OHDSI, 5 Years Later

## OHDSI Community Call
## Sept. 17 • 2024 • 11 am ET

# Upcoming Community Calls

| Date | Topic |
|---|---|
| Sept. 17 | The Book of OHDSI, Five Years Later |
| Sept. 24 | Recent OHDSI Publications |
| Oct. 1 | DARWIN EU® Review |
| Oct. 8 | TBA |
| Oct. 15 | Global Symposium Mad Minutes/Final Logistics |
| Oct. 22 | No Meeting due to Global Symposium |
| Oct. 29 | Welcome to OHDSI |

# Three Stages of The Journey

## Where Have We Been?

## Where Are We Now?

## Where Are We Going?

# OHDSI Shoutouts!

Congratulations to the team of **Jung-Joon Cha, Yunjin Yum, Yong Hyun Kim, Eung Ju Kim, Yoon Chan Rah, Euyhyun Park, Gi Jung Im, Jae-Jun Song, Sung-Won Chae, June Choi, and Hyung Joon Joo** on the publication of **Association of the protective effect of telmisartan on hearing loss among patients with hypertension** in *Frontiers in Neurology*.

frontiers | Frontiers in Neurology

TYPE Original Research
PUBLISHED 27 August 2024
DOI 10.3389/fneur.2024.1410389

Check for updates

OPEN ACCESS

EDITED BY
Leonard Rybak,
Southern Illinois University Carbondale, United States

REVIEWED BY
Takaomi Kurioka,
Kitasato University, Japan
Rodolfo Rothlin,
Independent Researcher, Buenos Aires, Argentina

*CORRESPONDENCE
June Choi
mednlaw@korea.ac.kr
Hyung Joon Joo
drjoohj@gmail.com

RECEIVED 14 April 2024
ACCEPTED 12 August 2024
PUBLISHED 27 August 2024

CITATION
Cha J-J, Yum Y, Kim YH, Kim EJ, Rah YC, Park E, Im GJ, Song J-J, Chae S-W, Choi J and Joo HJ (2024) Association of the protective effect of telmisartan on hearing loss among patients with hypertension. *Front. Neurol.* 15:1410389. doi: 10.3389/fneur.2024.1410389



## Association of the protective effect of telmisartan on hearing loss among patients with hypertension

Jung-Joon Cha[1], Yunjin Yum[2], Yong Hyun Kim[3], Eung Ju Kim[4], Yoon Chan Rah[5], Euyhyun Park[6], Gi Jung Im[6], Jae-Jun Song[7], Sung-Won Chae[7], June Choi[5]* and Hyung Joon Joo[1,8,9]*

[1]Division of Cardiology, Department of Internal Medicine, Korea University Anam Hospital, Seoul, Republic of Korea, [2]Department of Biostatistics, Korea University College of Medicine, Seoul, Republic of Korea, [3]Division of Cardiology, Department of Internal Medicine, Korea University Ansan Hospital, Ansan, Republic of Korea, [4]Division of Cardiology, Department of Internal Medicine, Korea University Guro Hospital, Seoul, Republic of Korea, [5]Department of Otorhinolaryngology-Head and Neck Surgery, Korea University Ansan Hospital, College of Medicine, Korea University, Ansan, Republic of Korea, [6]Department of Otorhinolaryngology-Head and Neck Surgery, Korea University Anam Hospital, College of Medicine, Korea University, Seoul, Republic of Korea, [7]Department of Otorhinolaryngology-Head and Neck Surgery, Korea University Guro Hospital, College of Medicine, Korea University, Seoul, Republic of Korea, [8]Department of Medical Informatics, Korea University College of Medicine, Seoul, Republic of Korea, [9]Korea University Research Institute for Medical Bigdata Science, College of Medicine, Korea University, Seoul, Republic of Korea

**Aim:** Hearing loss, affecting a significant portion of the global population, is prevented with peroxisome proliferator-activated receptor γ agonism. Understanding potential protective treatments is crucial for public health. We examine the effect of telmisartan, an antihypertensive drug and partial peroxisome proliferator-activated receptor γ agonist, on hearing loss in patients with hypertension.

**Method and results:** This retrospective cohort analysis used data from the OMOP Common Data Model database, encompassing information from three tertiary institutions in South Korea. The study included a substantial sample size of 860,103 people diagnosed with hypertension. The study included individuals who had been medically diagnosed with hypertension and had been prescribed antihypertensive drugs, including telmisartan. The study design was established to evaluate the comparative effects of telmisartan and other hypertension medications on hearing loss. We used propensity score matching (PSM) to create a balanced cohort, reducing potential biases between the telmisartan and non-telmisartan groups. From the initial 860,103 patients with hypertension,

# OHDSI Shoutouts!

Congratulations to the team of **Atsuhisa Sato, Daloha Rodriguez-Molina, Kanae Yoshikawa-Ryan, Satoshi Yamashita, Suguru Okami, Fangfang Liu, Alfredo Farjat, Nikolaus Oberprieler, Csaba Kovesdy, Keizo Kanasaki, and David Vizcaya** on the publication of **Early Clinical Experience of Finerenone in People with Chronic Kidney Disease and Type 2 Diabetes in Japan-A Multi-Cohort Study from the FOUNTAIN (FinerenOne mUltidatabase NeTwork for Evidence generAtIoN) Platform** in *Journals of Clinical Medicine*.

Journal of *Clinical Medicine*

MDPI

*Article*

## Early Clinical Experience of Finerenone in People with Chronic Kidney Disease and Type 2 Diabetes in Japan—A Multi-Cohort Study from the FOUNTAIN (FinerenOne mUltidatabase NeTwork for Evidence generAtIoN) Platform

**Atsuhisa Sato [1], Daloha Rodriguez-Molina [2], Kanae Yoshikawa-Ryan [3], Satoshi Yamashita [3], Suguru Okami [3,*], Fangfang Liu [2], Alfredo Farjat [2], Nikolaus G. Oberprieler [2], Csaba P. Kovesdy [4], Keizo Kanasaki [5,6] and David Vizcaya [2]**

1 Department of Nephrology and Hypertension, International University of Health and Welfare Shioya Hospital, Yaita 329-2145, Japan
2 Integrated Evidence Generation & Business Innovation, Bayer AG, 13342 Berlin, Germany; fangfang.liu@bayer.com (F.L.)
3 Medical Affairs & Pharmacovigilance, Bayer Yakuhin Ltd., Breeze Tower, 2-4-9 Umeda, Kita-ku, Osaka 530-0001, Japan; satoshi.yamashita@bayer.com (S.Y.)
4 Division of Nephrology, Department of Medicine, University of Tennessee Health Science Center, Memphis, TN 38163, USA
5 Department of Internal Medicine 1, Faculty of Medicine, Shimane University, 89-1 Enya-cho, Izumo 693-8501, Japan
6 Center for Integrated Kidney Research and Advance, Faculty of Medicine, Shimane University, 89-1 Enya-cho, Izumo 693-8501, Japan
* Correspondence: suguru.okami@bayer.com; Tel.: +81-6-6133-7000

check for updates

**Citation:** Sato, A.; Rodriguez-Molina, D.; Yoshikawa-Ryan, K.; Yamashita, S.; Okami, S.; Liu, F.; Farjat, A.; Oberprieler, N.G.; Kovesdy, C.P.; Kanasaki, K.; et al. Early Clinical Experience of Finerenone in People with Chronic Kidney Disease and Type 2 Diabetes in Japan—A Multi-Cohort Study from the FOUNTAIN (FinerenOne mUltidatabase NeTwork for Evidence generAtIoN) Platform. *J. Clin. Med.* **2024**, *13*, 5107. https://doi.org/10.3390/jcm13175107

Academic Editor: Abdallah Al-Salameh

Received: 12 July 2024
Revised: 15 August 2024
Accepted: 26 August 2024
Published: 28 August 2024

**Abstract: Background:** In the phase 3 clinical trials FIGARO-DKD and FIDELIO-DKD, finerenone reduced the risk of cardiovascular and kidney events among people with chronic kidney disease (CKD) and type 2 diabetes (T2D). Evidence regarding finerenone use in real-world settings is limited. **Methods:** A retrospective cohort study (NCT06278207) using two Japanese nationwide hospital-based databases provided by Medical Data Vision (MDV) and Real World Data Co., Ltd. (RWD Co., Kyoto Japan), converted to the OMOP common data model, was conducted. Persons with CKD and T2D initiating finerenone from 1 July 2021, to 30 August 2023, were included. Baseline characteristics were described. The occurrence of hyperkalemia after finerenone initiation was assessed. **Results:** 1029 new users of finerenone were included (967 from MDV and 62 from RWD Co.). Mean age was 69.5 and 72.4 years with 27.3% and 27.4% being female in the MDV and RWD Co. databases, respectively. Hypertension (92 and 95%), hyperlipidemia (59 and 71%), and congestive heart failure (60 and 66%) were commonly observed comorbidities. At baseline, 80% of persons were prescribed angiotensin-converting-enzyme inhibitors or angiotensin-receptor blockers. Sodium–glucose cotransporter 2 inhibitors and glucagon-like peptide 1 receptor agonists were prescribed in 72% and 30% of the study population, respectively. The incidence proportions of hyperkalemia were 2.16 and 2.70 per 100 persons in the MDV and RWD Co. databases, respectively. There were no hospitalizations associated with hyperkalemia observed in either of the two datasets. **Conclusions:**

# 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 | 1 pm | Common Data Model |
| Wednesday | 1 pm | Perinatal & Reproductive Health |
| Wednesday | 4 pm | Joint Vulcan/OHDSI Meeting |
| Wednesday | 7 pm | Medical Imaging |
| Thursday | 8 am | Medical Devices |
| Thursday | 8 am | OHDSI India Community Call |
| Thursday | 9 am | OMOP CDM Oncology Vocabulary/Development Subgroup |
| Thursday | 9:30 am | Themis |
| Thursday | 12 pm | HADES |
| Thursday | 6 pm | Eyecare and Vision Research |
| Thursday | 7 pm | Dentistry |
| Friday | 10 am | GIS-Geographic Information System |
| Friday | 10:30 am | Open-Source Community |
| Friday | 11:30 am | Steering Group |
| Monday | 9 am | Vaccine Vocabulary |
| Monday | 10 am | Healthcare Systems Interest Group |
| Tuesday | 9 am | OMOP CDM Oncology Genomic Subgroup |

# Congratulations, 2024 Titan Award nominees!

Alexander Davydov • Andrew Kanter • Anna Ostropolets • Anthony Sena • April Olympians Team • Asieh Golozar • Ben Martin • Benjamin Viernes • Christopher Mecoli • Cindy Cai • Clair Blacketer • Cynthia Sung • Daniel Morales • Danielle Boyce • DARWIN EU Development Team • Elisse Katzman • Evanette Burrows • Eye Care and Vision Research Workgroup • Frank DeFalco • George Hripcsak • Greg Klebanov • Henrik John • Hsin Yi Chen • J Swetha Kiranmayi • Jack Janetzki • James Weaver • Jared Houghtaling • Jen Park • Joel Swerdel • John Gresh • Jung Ho Kim • Justin Manjourides • Kyle Zollo-Venecek • Liesbet Peeters • Linying Zhang • Louis Hendricks • Maarten van Kessel • Manlik Kwong • Marc Suchard • Marta Pineda-Moncusi • Marti Catala Sabate • Martijn Schuemie • Martin Lavallee • Maxim Moinat • Michael Gurley • Michael Matheny • Michel Walravens • Michelle Hribar • Minnesota EHR Consortium Health Trends Across Communities Project Team • Montse Camprubi • Mengling ‘Mornin’ Feng • Natthawut 'Max' Adulyanukosol • OHDSI APAC ETL Team • OHDSI Standardized Vocabularies Team • Oleg Zhuk • Parthiban Sulur • Polina Talapova • Qi Yang • Renske Los • Rich Boyce • Robert Koski • Robert Miller • Roger Carlson • Scott DuVall • Thamir Alshammary • Theresa Burkard • Thomas Falconer • Tom Seinen • Vishnu Chandrabalan • Vlad Korsik • Will Kelly • Zhen Lin

# 2024 APAC Symposium

## Dec. 4-8 • Marina Bay Sands & National University of Singapore (NUS)

## Registration is OPEN!

### Preliminary Dates To Know

Oct. 6: Collaborator Showcase Submission Deadline
Oct. 7-24: Collaborator Showcase Submission Review
Oct. 31: Notification of Acceptance

### Symposium Agenda

Dec. 4: Tutorial at NUS
Dec. 5-6: Main Conference at Marina Bay Sands
Dec. 7-8: Datathon at NUS

**ohdsi.org/APAC2024**

# 2024 India Symposium

## Oct. 5 • Jio World Convention Centre • Mumbai

# 2024 Global Symposium

## Oct. 22-24 • Hyatt Regency Hotel • New Brunswick • N.J.

Registration is OPEN for the 2024 OHDSI Global Symposium. Collaborator Showcase notifications are taking place this week. Agendas and tutorial/workgroup schedules are posted.

**Tuesday:** Tutorials
**Wednesday:** Plenary/Showcase
**Thursday:** Workgroup Activities


**ohdsi.org/OHDSI2024**

# The Center for Advanced Healthcare Research Informatics (CAHRI) at Tufts Medicine welcomes:

**Cavin Ward-Caviness, PhD**

*Senior Computational Biologist in the Public Health and Integrated Toxicology Division of the US Environmental Protection Agency*

**'Successes and Lessons Learned from Integrating Environmental Data into Diverse EHR Resources'**

September 26, 2024, 11am-12pm EST
Virtually via Zoom

Please contact Marty Alvarez at malvarez2@tuftsmedicalcenter.org for calendar invite or questions.

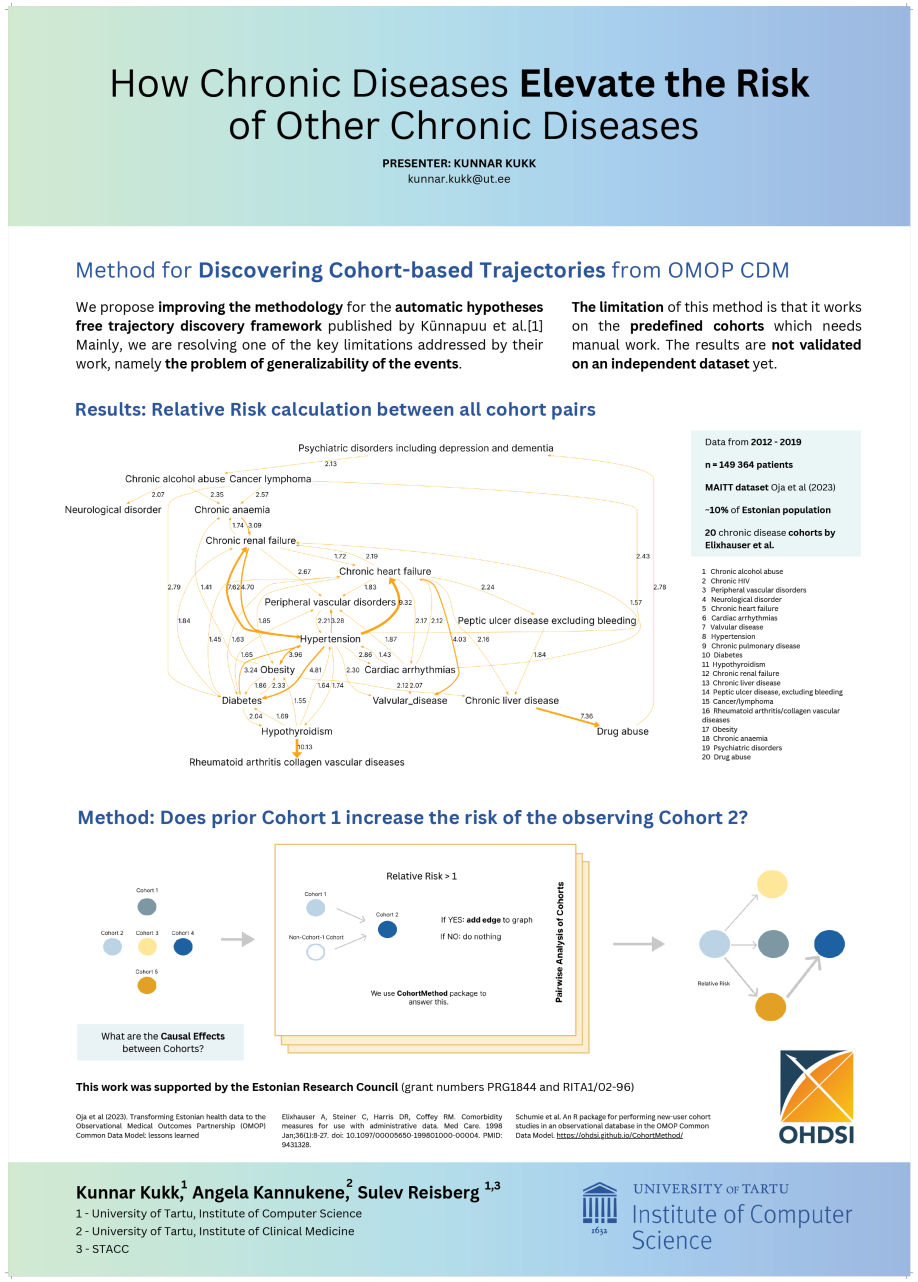
TuftsMedicine
Tufts Medical Center

# #OHDSISocialShowcase

## MONDAY

# How Chronic Diseases Elevate the Risk of Other Chronic Diseases

(Kunnar Kukk, Angela Kannukene, Sulev Reisberg)



# #OHDSISocialShowcase

## TUESDAY

# Code reuse in OHDSI standardized vocabularies: scope of the problem and methods to solve

(Dmitry Dymshyts, Clair Blacketer, Evanette K Burrows, Anna Ostropolets, Erica A Voss)

## HCPCS and NDC codes change meaning over time. Solving it in OHDSI vocabulary and ETL

*Title: Code reuse in OHDSI standardized vocabularies: scope of the problem and methods to solve*

**Background: 43 HCPCS and 82 NDC codes that changed their meaning in period from 2017 until now were identified. Their absolute count is small, but they contain novel therapies (see examples below)**

Table 1. Examples of reused codes.

| code | vocabulary | previous meaning | old meaning end date | current meaning start date | current meaning |
|---|---|---|---|---|---|
| J9350 | HCPCS | Injection, **topotecan**, 4 mg (Deprecated) | 2014-11-11 | 2023-07-01 | Injection, **mosunetuzumab**-axgb, 1 mg |
| J1440 | HCPCS | Injection, **filgrastim** (g-csf), 300 mcg (Deprecated) | 2013-12-31 | 2023-07-01 | **Fecal microbiota**, live - jslm, 1 ml |
| J9380 | HCPCS | **Vincristine** sulfate, 5 mg (Deprecated) | 2014-11-11 | 2023-07-01 | Injection, **teclistamab**-cqyv, 0.5 mg |
| G0327 | HCPCS | **END STAGE RENAL DISEASE** (ESRD) RELATED SERVICES LESS THAN FULL MONTH, PER DAY; ... | 2013-08-28 | 2021-07-01 | **Colorectal cancer screening**; blood-based biomarker |
| Q2040 | HCPCS | Botulinum Toxin Type A Injection | 2011-12-31 | 2018-01-01 | **Tisagenlecleucel**, up to 250 million car-positive viable t cells... |
| Q2041 | HCPCS | von Willebrand factor 1 UNT Injectable Solution | 2011-12-31 | 2019-01-01 | **Axicabtagene** ciloleucel, up to 200 million autologous anti-cd19 car positive viable t cells... |
| Q2042 | HCPCS | 17-alpha-Hydroxyprogesterone Injection | 2011-12-31 | 2019-01-01 | Tisagenlecleucel, up to 600 million car-positive viable t cells... |
| 516550 86026 | NDC | Alprazolam 0.5 MG Oral Tablet | 2013-03-01 | 2023-08-01 | estradiol 0.5 MG Oral Tablet |
| 516550 19326 | NDC | Furosemide 20 MG Oral Tablet | 2013-03-01 | 2023-04-01 | 24 HR metoprolol succinate 100 MG Extended Release Oral Tablet |

**Note, the meaning shown in OHDSI vocabulary is in GREEN.**

Currently several concepts with the same vocabulary_id and code aren't allowed, that's why **either current or historical meaning is reflected, but never both**

Proposed vocabulary solution:

Table 2. Proposed concept table:

| concept_code | Concept_name | valid_start_date | valid_end_date | invalid_reason* | concept_id |
|---|---|---|---|---|---|
| J9380 | Vincristine sulfate, 5 mg | 1970-01-01 | 2014-11-11 | R | 1234 |
| J9380 | Injection, teclistamab-cqyv, 0.5 mg | 2023-07-01 | 2099-12-31 | R | 5678 |

**Each concept is treated separately in the other vocabulary tables having its own mappings, hierarchy (if standard) and synonyms**

*Invalid_reason = 'R' means that concept is "reused", and requires special treatment by the ETL matching not only source codes and vocabulary, but event dates and concept valid_dates

Figure 1. Current CDM: drug exposure – same drug all the time


Figure 2. Resulting CDM: drug exposure – different drugs at different times


Figure 3. How to define Teclistimab exposure currently

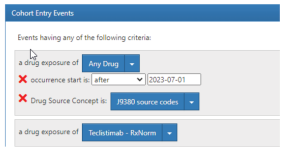
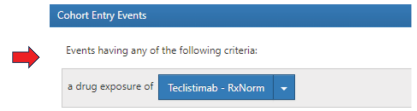



Figure 4. How to define Teclistimab exposure with applied changes




**Conclusion:** OHDSI solves the HCPCS and NDC code reuse problem on the vocabulary / CDM level by mapping events from different time periods to different concepts, so researchers can address them directly, without specifying dates.

Dmitry Dymshyts, Clair Blacketer,
Evanette K Burrows, Anna Ostropolets, Erica A Voss

Johnson & Johnson

# #OHDSISocialShowcase

## WEDNESDAY

# Dimension Reduction Techiques for Clinical Predictions Models on Health Care Data

(Roëlle Bänffer, Aniek Markus, Tom Seinen)

## DIMENSION REDUCTION TECHNIQUES FOR CLINICAL PREDICTION MODELS ON HEALTH CARE DATA

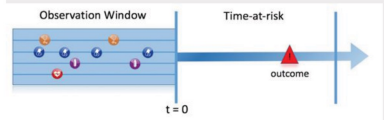
*Research aim: What is the impact of applying different types of dimension reduction (DR) techniques on the predictive performance, generalizability, and interpretability of clinical prediction models?*

**Problem formalization**

- The Electronic Health Record(EHR) data is **complex** and contains many medical codes for each diagnosis, procedure, and medication of a patient.
- The **sparsity** and **high dimensionality** of these EHRs present significant challenges in implementing widely applicable clinical prediction models.
- Current existing methods make use of techniques like **feature selection** which leads to a **loss of information**.

*Many clinical events exist (high dimensionality), but patients typically only experience a few (sparsity)*

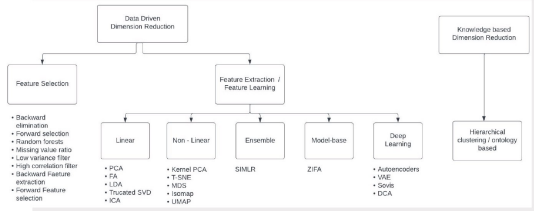
**Patient Level Prediction framework**




**Prediction problem:**
Among a population at risk, we aim to predict which patients at a defined moment in time (t=0) will experience some outcome during a time-at-risk. Prediction is done using only information about the patients in an observation window prior to that moment in time.

Current prediction question: Among hospital discharges of adult(18+) patients, who will go on to have hospital admission with 2 to 30 days?
Observation window: 365 days

**Data Driven versus Knowledge Based**

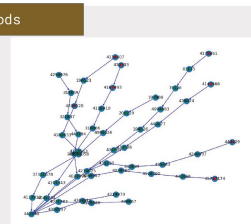


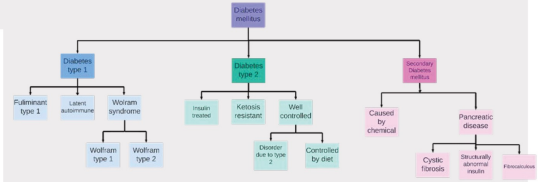

**Data Driven Dimension Reduction:**
A data driven algorithm learns the data structure and the characteristics of the data components, denoises the data, increases the separation between the components and projects the data onto a lower number of dimensions.

**Knowledge based Dimension Reduction:**
Knowledge based makes use of the hierarchical structure of the OMOP standarized vocabulary to reduce the number of dimension. The hierarchical structure can be verified by an domain expert.

**Methods**

*Example of Hierarchical structure created from data*






*Example of structure used for Knowledge based Dimension Reduction*

Table 4.2: Knowledge based approach (Levels based)

| Input | Number of Candidate features | Model AUC train | Model AUC test | Number of Model features |
|---|---|---|---|---|
| L = 1 | 4 | 52.60 | 63.00 | 3 |
| L = 2 | 33 | 51.12 | 55.33 | 4 |
| L = 3 | 1187 | 77.22 | 73.97 | 68 |
| L = 4 | 2008 | 76.85 | 76.21 | 85 |
| L = 5 | 2206 | 76.03 | 69.31 | 112 |
| L = 6 | 1403 | 69.86 | 58.27 | 77 |

Table 4.3: Knowledge based versus Data driven approach

| Input | Number of Candidate features | Model AUC train | Model AUC test | Number of Model features |
|---|---|---|---|---|
| Condition Occurence | 2795 | 77.51 | 75.37 | 85 |
| Condition Group | 2633 | 78.67 | 75.38 | 121 |
| PCA | 100 | 49.29 | 63.39 | 1 |
| SVD | 100 | 74.18 | 52.86 | 1 |
| Auto Encoder | 7 | 52.84 | 51.00 | 2 |

Health Data Science
Roëlle Bänffer, Aniek Markus, Tom Seinen
Erasmus MC

OHDSI

# #OHDSISocialShowcase

## THURSDAY

## PHEMS: validating novel federated ecosystems for analytics and synthetic data generation methods through real-world data investigations, particularly in the context of pediatric healthcare

(Sofia Bazakou, Lydia Briggs, Marinel Cavelaars, Roger Domingo Espinos, Aida Felipe Villalobos, Katariina Gehrmann, Liam Glueck, Iolanda Jordan Garcia, Jan Willem Kuiper, Jennifer McIntosh, Cristina Ruiz Herguido, Andrew Taylor, Marja Vaitinen, Arnau Valls Esteve, Gary Zhen Yuan Liew, Azadeh Tafreshiha, Guus Wilmink)

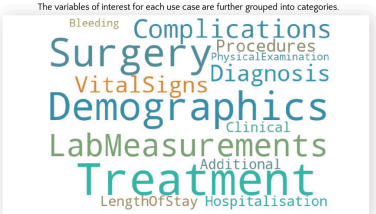
PHEMS: **validating novel federated ecosystems** for analytics and **synthetic data generation** methods through real-world data investigations, particularly in the context of **pediatric healthcare**

*New strategies in Health Data Sharing – Clinical Use Cases*

***Pediatric Hospitals as European drivers for multi-party computation and synthetic data generation capabilities across clinical specialties and data types (PHEMS, an EU-funded project) aims to establish an open and decentralized health data ecosystem for accessing health data across multiple European hospitals. Data federation (federated analytics and learning) will be applied to overcome significant obstacles in cross-border collaboration while complying with the European Union's General Data Protection Regulations (GDPR) and AI Act. Additionally, PHEMS will develop an innovative data synthesis and anonymization pipeline for use in the federated ecosystem for rare disease research.***

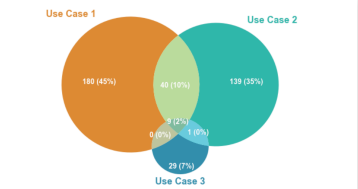
Variable categories

The variables of interest for each use case are further grouped into categories.




Variable overlap

An overview of the number of variables to be collected for each use case and their overlap.




**1 Cardiology operations benchmarking**

Supporting **creation of benchmarking standard** and **promote a culture of benchmarking** across pediatric cardiac institutions, enabling **the adoption of 'best-practice' across institutions**

Led by Great Ormond Street Hospital for Children (GOSH)

**2 Pediatric Intensive Care Unit Sepsis**

Investigating the benefits of the federated ecosystem to **develop, train and test algorithms to predict sepsis** on a large scale between pediatric intensive care units in four large European children's hospitals

Led by Sant Joan de Déu Barcelona Children's Hospital (HSJD)

**3 Hematology - hemophilia**

**Developing and testing a machine learning-based prediction algorithm** to improve treatment for pediatric patients with **hemophilia A or B**

Led by Erasmus University Medical Centre Rotterdam (Erasmus)

**At the current state of the project, the achievement worth mentioning is the curation of the variables essential to the use case studies. This preliminary work, in the first months of the consortium, lays the foundation for the OMOP data standardization at the hospitals and the synthetic data generation.**

OHDSI

Sofia Bazakou, Lydia Briggs, Marinel Cavelaars, Roger Domingo Espinos, Aida Felipe Villalobos, Katariina Gehrmann, Liam Glueck, Iolanda Jordan Garcia, Jan Willem Kuiper, Jennifer McIntosh, Cristina Ruiz Herguido, Andrew Taylor, Marja Vaitinen, Arnau Valls Esteve, Gary Zhen Yuan Liew, Azadeh Tafreshiha, Guus Wilmink

phems | Funded by the European Union | UK Research and Innovation | The Hyve | Erasmus MC | HUS | SJD Sant Joan de Déu Fundació de Recerca | NHS Great Ormond Street Hospital for Children | ECHO | SJD Sant Joan de Déu Barcelona Hospital

# #OHDSISocialShowcase

## FRIDAY

# Custom vocabulary techniques in ETL to OMOP CDM

(Tatsiana Skuhareuskaya, Vlad Korsik, Vojtech Huser, Alexander Davydov)

## Custom vocabulary techniques in ETL to OMOP CDM

*An overview of aspects to consider when creating and maintaining a custom vocabulary*

**Background:** While OHDSI Standardized vocabularies available through Athena (e.g. ICD10CM, NDC) greatly facilitate the mapping of a big portion of source data, other source data encoded in terminologies not included in Athena or free text requires many special considerations. Data such as Medical histories, Allergies and Registries, including Surveys, are valuable for research purposes and need to be captured in the CDM. Currently creating a custom vocabulary with concept_id of >2 billion is a default approach (Table 1). Here we describe the reasoning behind some of the crucial choices made when creating a custom vocabulary - identification of custom concepts using pre- and post-coordination. We also touch on domain drift as an important factor when maintaining a custom vocabulary or developing and supporting custom terminologies.

| Parameter | STCM | C/CR |
|---|---|---|
| General description | Creation of a mapping from source_code to a standard_concept_id | Creation of a custom concept (with concept_id>2bil) and mapping it to a standard_concept_id |
| Loading | Easy: table is designed for insertion of records | Requires alteration of vocabulary tables from Athena |
| Destiny of unmapped codes | Live only as source_codes | Live as custom concepts without mapping, i.e. as event_source_concept_id and can be queried by Atlas |
| Maps_to_value relationships | Does not support | Supports |
| ETL logic | Requires additional querying of the STCM | Allows for a usual lookup of a concept in the Concept table |
| Research use | None: data is not visible in Atlas or the Concept table | Custom concepts can be used in network studies through Atlas |
| **Conclusion** | **Easy to implement, but not sharable** | **More complex, but sharable and usable in network research** |

Table 1: Comparison of source_to_concept_map and concept/concept_relationship approaches to ETL

**Methods:** The results are based on our experience of converting 34 unique datasets and their variations with all types of US- and ex-US data sources: EHR (e.g. Flatiron, Epic), claims (e.g. JMDC, CPRD-family), registry (e.g. NAACCR, disease- or study-specific), SDTM, etc. In our experience, the number of newly created custom vocabularies ranged between 4 to 57 per dataset (with a median of 16).

**Results:** When starting work on a vocabulary and choosing entities to **define** source_codes, these are some of the most crucial points:

- Domain integrity among codes in the same vocabulary
- Creating codes as event-value pairs (i.e. post-coordinate) vs defined facts (i.e. pre-coordinate) (Figure 1)
- What entities would enrich the definition (units, dosages/concentrations, attributes)
- What are the desired implications in the phenotype construction process?

When the vocabulary is compiled, QA is the next step consisting of three parts:

**Structural integrity checks, e.g.:**
- Are concept codes unique within the custom vocabulary?
- Is the structure compliant with DDL and the content is referenced to the concept table?

**Semantic integrity checks, e.g.:**
- Are mappings to Standard concepts correct?

**Content integrity checks, e.g.:**
- Do the concept domains and concept classes of target concepts comply with the CDM conventions?
- Are the post-coordinated (e.g. 1-to-many, event-value pairs) mappings logically justified and structurally supported?

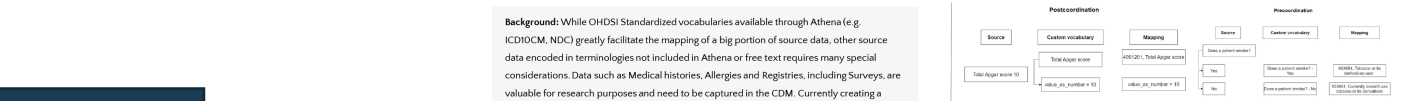



Figure 1: Examples of pre-coordination and post-coordination when creating a custom vocabulary (left) and changing mappings (right)

Afterwards, the iterative process of **maintaining** the custom vocabulary takes place. When OMOP Vocabularies are updated, target concepts can change their attributes and live cycle (Standard to non-Standard; Domain change is also possible) (Figure 2) and remappings may be necessary. Post-coordination should be taken into account and 'Maps to value' relationships should be created where necessary (Figure 3). This creates a need for additional semantic integrity checks, especially for cases where post-coordination has previously been used.

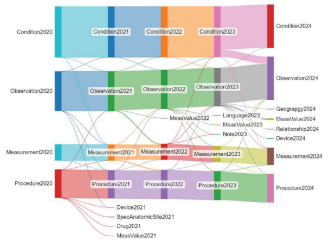



Figure 2: Domain changes for Standard concepts used in ETL across 2020-2024 years

**Discussion:** To fully enable the benefit of the Common Data Model and portable analysis across datasets, semantic standardization (usage of a thoughtfully selected set of Standard concepts as targets) is equally important and intertwined with structural standardization (column and table structure).
Emerging use cases and the overall complexity of the tasks require for the development of new approaches to custom vocabulary construction. Examples of such new approaches are:

- Wide mapping table [1] - alternative mapping table format to allow for multiple target entities
- Logic groups [1] - addition to the concept_relationship table which connects 'Maps to' relationships with their respective attributes
- Value_source_concept_id field - approach for better standartization of survey data.
- Common Data Environment [2] - approach for mapping harmonization and enrichment

**Limitations:** OMOP vocabularies, as well as the CDM itself, are always the work in progress. In order to benefit from new Vocabulary, ETL or CDM developments, it is crucial to constantly re-evaluate and adjust the existing approaches and mappings.

**References:** [1] [2]

Tatsiana Skuhareuskaya, Vlad Korsik, Vojtech Huser, Alexander Davydov

# Where Are We Going?

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

Please feel free to promote your #OHDSI2024 workshop or workgroup activity!

# Three Stages of The Journey

## Where Have We Been?

## Where Are We Now?

## Where Are We Going?

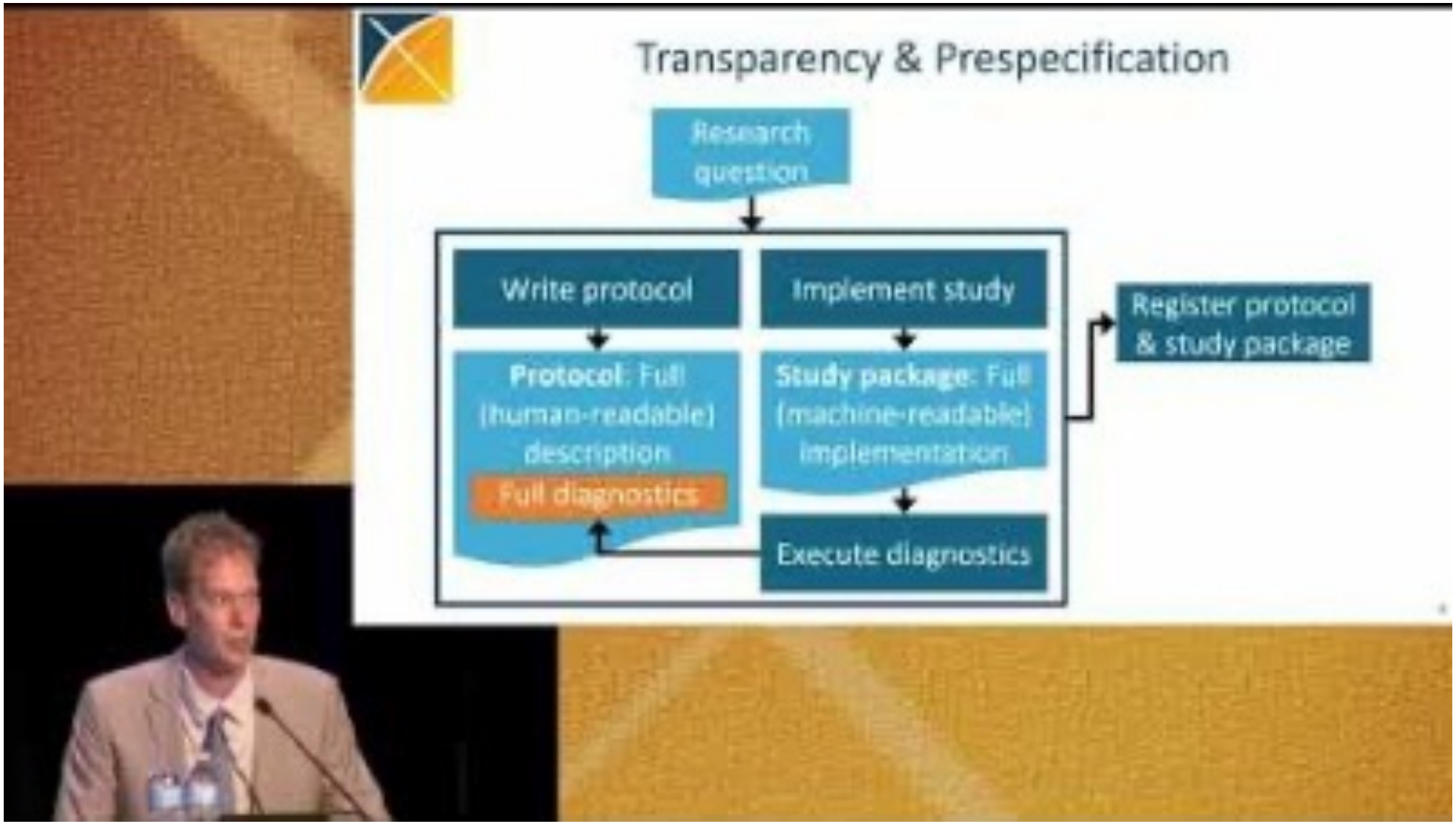
# Three Stages of The Journey

## Where Have We Been?

## Where Were We Five Years Ago?

## Where Are We Going?

# Five Years Ago



# The Team

## Thank You To Our Book of OHDSI Contributors

| | | | | |
|---|---|---|---|---|
| Hamed Abedtash | Mustafa Ascha | Mark Beno | Clair Blacketer | David Blatt |
| Brian Christian | Gino Cloft | Frank DeFalco | Sara Dempster | Jon Duke |
| Sergio Eslava | Clark Evans | Thomas Falconer | George Hripscak | Vojtech Huser |
| Mark Khayter | Greg Klebanov | Kristin Kostka | Bob Lanese | Wanda Lattimore |
| Chun Li | David Madigan | Sindhoosha Malay | Harry Menegay | Akihiko Nishimura |
| Ellen Palmer | Nirav Patil | Jose Posada | Nicole Pratt | Dani Prieto-Alhambra |
| Christian Reich | Jenna Reps | Peter Rijnbeek | Patrick Ryan | Craig Sachson |
| Izzy Saridakis | Paola Saroufim | Martijn Schuemie | Sarah Seager | Anthony Sena |
| Sunah Song | Matthew Spotnitz | Marc Suchard | Joel Swerdel | Devin Tian |
| Don Torok | Kees van Bochove | Mui Van Zandt | Erica Voss | Kristin Waite |
| Mike Warfe | Jamie Weaver | James Wiggins | Andrew Williams | Seng Chan You |

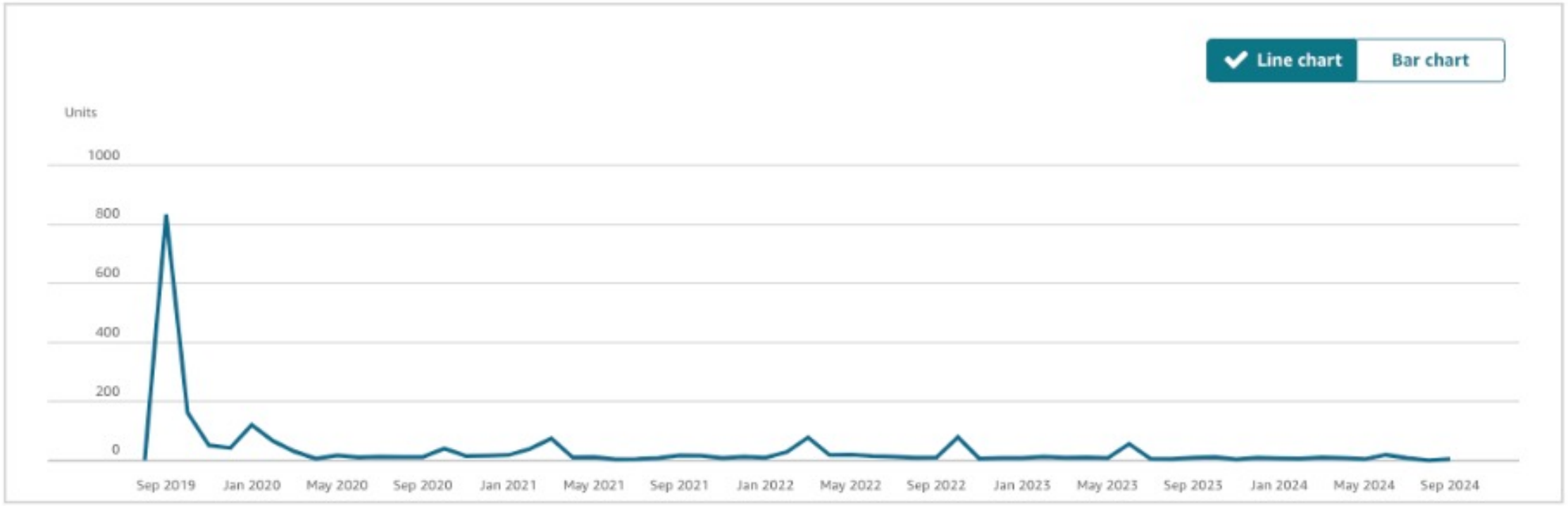
# Over The Last Five Years

Units Processed

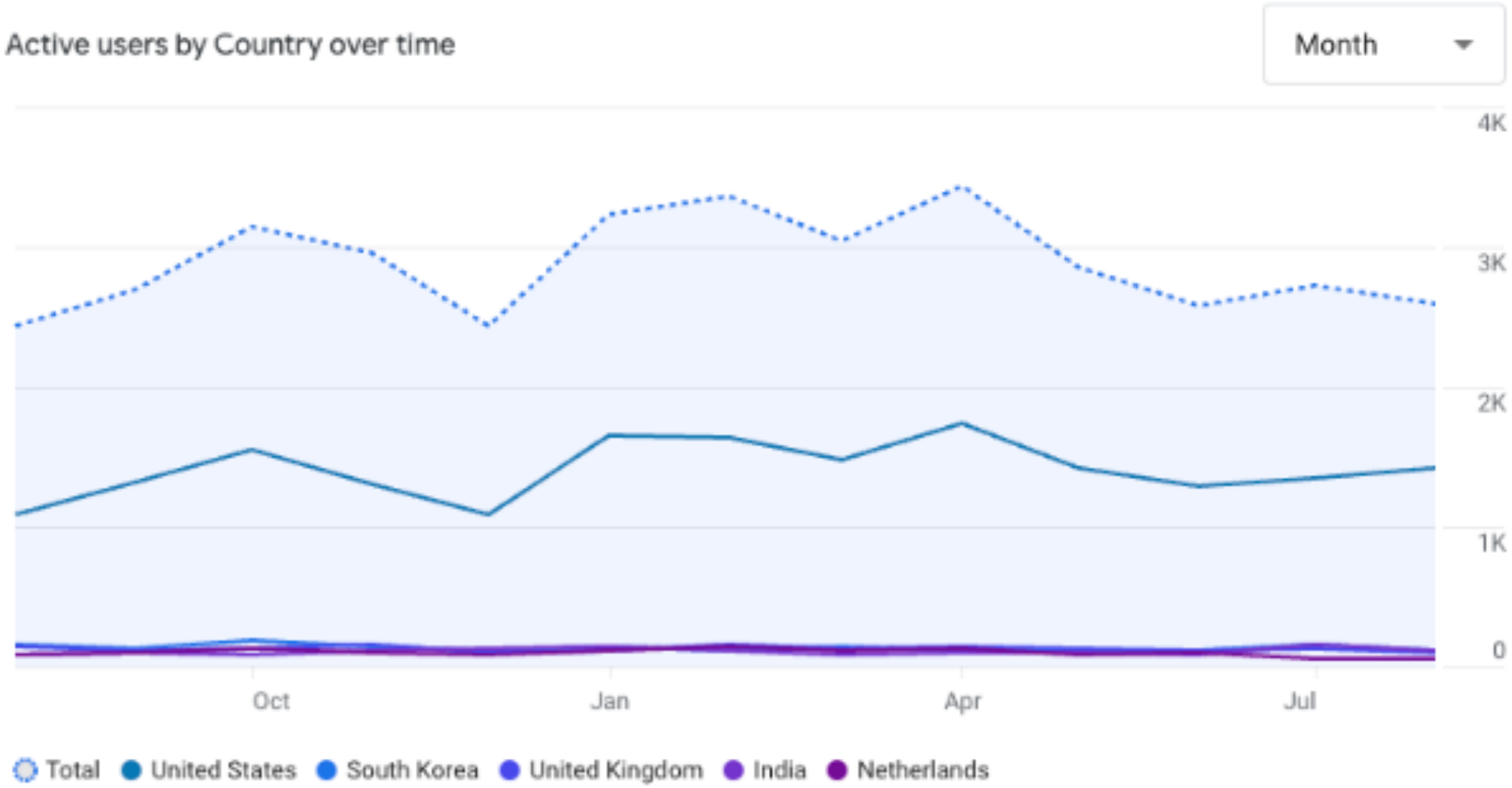
2,172

All 1 books

Aug 30, 2019 - Sep 13, 2024



# Over The Last Five Years




# 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