CDM Update Process

OHDSI Community Call
July 26, 2022 • 11 am ET

Upcoming OHDSI Community Calls

Date	Topic
Aug. 2	Building A Community Within Your Organization
Aug. 9	Around The Asia-Pacific (APAC) Community
Aug. 16	OHDSI "Speed Dating"
Aug. 23	Workgroup Updates
Aug. 30	EHDEN Academy/EHDEN Portal

Upcoming OHDSI Community Calls

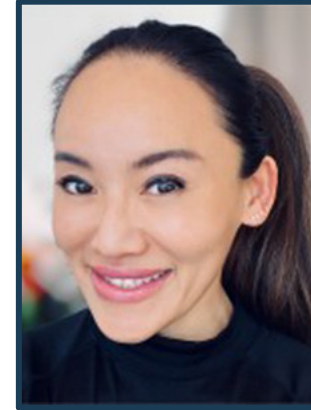| Date	Topic
Aug. 2	Building A Community Within Your Organization
Aug. 9	Around The Asia-Pacific (APAC) Community
Aug. 16	OHDSI "Speed Dating"
Aug. 23	Workgroup Updates
Aug. 30	EHDEN Academy/EHDEN Portal

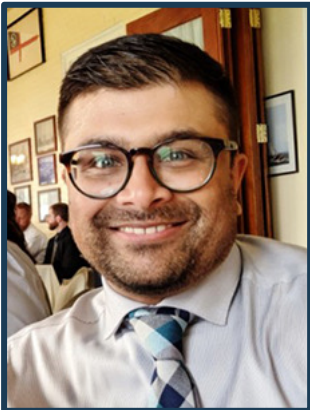# Aug. 2 Community Call: Building Organizational Support Within Your Community

Greg Klebanov

CTO/SVP •
Odysseus Data Services,
Inc.

Keran Moll

Director, HEOR Real World
Data & Analytics Research
• Regeneron

Ajit Londhe

Senior Manager, Center for
Observational Research •
AMGEN

Paul Nagy

Program Director for
Graduate Training in
Biomedical Informatics and
Data Science • Johns
Hopkins University

OHDSI Shoutouts! 👏

Congratulations to the team of **Jimmy Phuong, Stephanie Hong, Matvey B. Palchuk, Juan Espinoza, Daniella Meeker, David A. Dorr, Galina Lozinski, Charisse Madlock-Brown, and William G. Adams** on the publication of **Advancing Interoperability of Patient-level Social Determinants of Health Data to Support COVID-19 Research** in the AMIA Annual Symposium Proceedings Archive.

Advancing Interoperability of Patient-level Social Determinants of Health Data to Support COVID-19 Research

Jimmy Phuong, [1,2] Stephanie Hong, BS, [3] Matvey B. Palchuk, [4] Juan Espinoza, [5] Daniella Meeker, [6] David A. Dorr, [7] Galina Lozinski, [8] Charisse Madlock-Brown, [9] and William G. Adams [8]

▶ Author information ▶ Copyright and License information   Disclaimer

Abstract

Go to: ▶

Including social determinants of health (SDoH) data in health outcomes research is essential for studying the sources of healthcare disparities and developing strategies to mitigate stressors. In this report, we describe a pragmatic design and approach to explore the encoding needs for transmitting SDoH screening tool responses from a large safety-net hospital into the National Covid Cohort Collaborative (N3C) OMOP dataset. We provide a stepwise account of designing data mapping and ingestion for patient-level SDoH and summarize the results of screening. Our approach demonstrates that sharing of these important data - typically stored as non-standard, EHR vendor specific codes - is feasible. As SDoH screening gains broader use nationally, the approach described in this paper could be used for other screening instruments and improve the interoperability of these important data.

OHDSI Shoutouts!

Congratulations to the team of **Martijn Schuemie, Faaizah Arshad, Nicole Pratt, Fredrik Nyberg, Thamir Alshammari, George Hripcsak, Patrick Ryan, Daniel Prieto-Alhambra, Lana Lai, Xintong Li, Stephen Fortin, Evan Minty and Marc Suchard** on the publication of **Vaccine Safety Surveillance Using Routinely Collected Healthcare Data—An Empirical Evaluation of Epidemiological Designs** in Frontiers in Pharmacology.

Vaccine Safety Surveillance Using Routinely Collected Healthcare Data —An Empirical Evaluation of Epidemiological Designs

Martijn J. Schuemie [1,2,3]*, Faaizah Arshad [1,3], Nicole Pratt [4], Fredrik Nyberg [5], Thamir M Alshammari [6], George Hripcsak [1,7], Patrick Ryan [1,2,7], Daniel Prieto-Alhambra [8,9], Lana Y. H. Lai [10], Xintong Li [11], Stephen Fortin [2], Evan Minty [10] and Marc A. Suchard [1,3,12]

[1]Observational Health Data Sciences and Informatics, New York, NY, United States, [2]Observational Health Data Analytics, Janssen R&D, Titusville, NJ, United States, [3]Department of Biostatistics, University of California, Los Angeles, Los Angeles, CA, United States, [4]Quality Use of Medicines and Pharmacy Research Centre, Clinical and Health Sciences, University of South Australia, Adelaide, SA, Australia, [5]School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden, [6]College of Pharmacy, Prince Sattam Bin Abdulaziz University, Riyadh, Saudi Arabia, [7]Department of Biomedical Informatics, Columbia University, New York, NY, United States, [8]Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, United Kingdom, [9]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherlands, [10]O'Brien Institute for Public Health, Faculty of Medicine, University of Calgary, Calgary, AB, Canada, [11]Division of Medical Sciences, University of Manchester, Manchester, United Kingdom, [12]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, United States

Background: Routinely collected healthcare data such as administrative claims and electronic health records (EHR) can complement clinical trials and spontaneous reports to detect previously unknown risks of vaccines, but uncertainty remains about the behavior of alternative epidemiologic designs to detect and declare a true risk early.

Methods: Using three claims and one EHR database, we evaluate several variants of the case-control, comparative cohort, historical comparator, and self-controlled designs against historical vaccinations using real negative control outcomes (outcomes with no evidence to suggest that they could be caused by the vaccines) and simulated positive control outcomes.

OHDSI Shoutouts!

Any shoutouts from the community? Please share and help promote and celebrate OHDSI work!

Have a study published? Please send to sachson@ohdsi.org so we can share during this call and on our social channels.
Let's work together to promote the collaborative work happening in OHDSI!

Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Tuesday	3 pm	OMOP CDM Oncology Outreach/Research Subgroup
Wednesday	11 am	Open-Source Community
Wednesday	12 pm	Latin America
Wednesday	12 pm	FHIR and OMOP Terminologies Subgroup (ZOOM)
Wednesday	7 pm	Medical Imaging
Thursday	10 am	Data Quality Dashboard
Thursday	12 pm	FHIR and OMOP Oncology Subgroup
Friday	9 am	GIS-Geographic Information System
Monday	10 am	Healthcare Special Interest Group
Tuesday	10 am	Common Data Model

www.ohdsi.org/upcoming-working-group-calls

OHDSI APAC Community Calls

The next Asia-Pacific (APAC) community call takes place Thursday, July 28 (July 27 in the Western Hemisphere) and will focus on two of the ongoing APAC network studies: **Comparison of mortality, morbidities & healthcare resources utilization between patients with and without a diagnosis of COVID-19**, and **Real world safety of treatments for multiple sclerosis**.

Date	Topic
July 14	APAC Study Quarterly Updates, Part 1
July 28	APAC Study Quarterly Updates, Part 2
Aug. 11	Working Group Updates #3
Aug. 25	Working Group Updates #4
Sept. 8	EU Chapter Sharing Session, Part 1
Sept. 22	EU Chapter Sharing Session, Part 2

ohdsi.org/APAC

OHDSI APAC Symposium

2022 OHDSI APAC Symposium Overview

November 12-13, 2022

Hosted in Taiwan by Taipei Medical University

ohdsi.org/APAC

Titan Awards Nominations Are Open

To recognize OHDSI collaborators (or collaborating institutions) for their contributions towards OHDSI's mission, the OHDSI Titan Awards were introduced at the 2018 Symposium and have been handed out at the U.S./Global Symposium each year since. Annually, community members are invited to nominate individuals or institutions they feel have made significant contributions towards advancing OHDSI's mission, vision and values. Once nominations are submitted, the OHDSI Titan Award Committee will select the award winners. Award winners will be announced before the networking reception at the annual symposium. The award categories, as well as all previous recipients, can be found below.

2021 OHDSI Titan Awards — Maxim Moinat, Data Standards
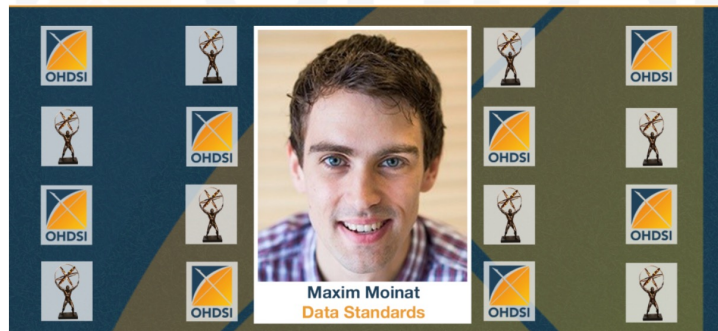
Titan Award for Data Standards – to recognize extraordinary contributions by an individual, organization, or team in development or evaluation in community data standards, including OMOP common data model and standardized vocabularies
- **2021 –** Maxim Moinat, The Hyve/Erasmus University Medical Center
- **2020 –** Clair Blacketer, Janssen Research and Development
- **2019 –** Oncology Workgroup (Michael Gurley, Northwestern Univ.; Rimma Belenkaya, Memorial Sloan Kettering Cancer Center; Robert Miller, Tufts CTSI)
- **2018 –** Vocabulary team (Christian Reich, IQVIA; Anna Ostropolets, Columbia Univ.; Dmitry Dymshyts, Odysseus Data Services)

2021 OHDSI Titan Awards — Yong Chen, Methodological Research

2021 OHDSI Titan Awards — Adam Black, Open-Source Development

2021 OHDSI Titan Awards — Asieh Golozar, Clinical Applications

2021 OHDSI Titan Awards — Erica Voss, Community Collaboration

2021 OHDSI Titan Awards — Mui Van Zandt, Community Leadership

2021 OHDSI Titan Awards — Faaizah Arshad, Community Support; Ross Williams, Community Support
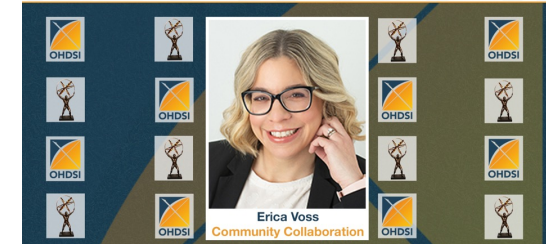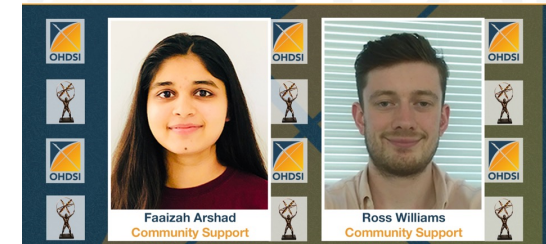
ohdsi.org/titan-awards

Job Openings

Peter Rijnbeek and his team at Erasmus University is hiring a Secretary for the Darwin EU Coordination Center and Department of Medical Informatics.

This position will be responsible for the day-to-day administrative tasks as the personal assistant for Peter Rijnbeek, and will also work as senior secretary for the Department of Medical Informatics.

The application deadline is Aug. 14.

Erasmus MC

Jobs   About Erasmus MC ⌄   Research facilities ⌄   Career ⌄   NL / EN

Secretary for the Darwin EU Coordination Center and Department of Medical Informatics

🕐 Closing date 14-08-2022   # 29.30.22.TG

🕐 36 hours

🎓 University of Applied Sciences

🧰 Administrative and Secretarial

Apply ▶

Questions? Contact our recruiter

☰ Job description   👥 Work environme...   👤 Qualifications a...   👤 Terms of employ...   ❓ More information

Job description

We are offering a unique opportunity to have a leading role in the setup and operation of the secretariat of the DARWIN EU® Coordination Center . In this high exposure initiative, the Department of Medical Informatics is operating a unique network of databases in Europe to answer research question from the European

Job Openings

Assistant professor **Brianne Oliveri-Mui** announced an opening for an Postdoctoral Fellow to work at the Roux Institute at Northeastern University.

If you are interested, please reach out to Dr. Mui at b.mui@northeastern.edu.

The link and more information will be available on the community calls page.

Observational Health Data Sciences and Informatics Postdoctoral Fellow

Apply

📍 Portland, ME

💼 Full time

🕐 Posted 30+ Days Ago

📄 R105484

About the Opportunity

The Roux Institute at Northeastern University has one opening for a Postdoctoral Research Fellow beginning on or about September 1, 2022. The fellow will have an opportunity to conduct observational and administrative database research (e.g., analysis of existing datasets) on health outcomes for older adults with HIV or LGBT older adults, under the supervision of the PI. The fellow will devote most of their time to independent research aligned with the PI's interests and across federated and local research models.

Position offers exceptional opportunity for collaboration at the OHDSI center on major projects in the U.S. and overseas. This research will directly improve our ability to use real world data to characterize under-represented and marginalized patient populations, construct population level estimates relating exposures to health outcomes, and to enhance clinical decision making through improved patient-level predictions. The term of fellowship appointment will be for two years, contingent on continued funding. Stipend will be commensurate with experience, based on levels mandated by NIH.

The main research areas specific to older people with HIV or in the LGBTQ+ communities are as follows:

- Measurement of comorbidities, care quality, health outcomes and healthcare utilization patterns
- Risk assessment of multimorbidity, healthcare and prescription access

Job Openings

Professor **Dani Prieto-Alhambra** and his team at the University of Oxford will be hiring two Research Assistants in Health Data Sciences.

The application deadline is August 8, 2022.

The link and more information will be available on the community calls page.

2022 OHDSI Symposium

Registration is OPEN for #OHDSI2022!

The 2022 OHDSI Symposium will be held Oct. 14-16 at the Bethesda North Marriott Hotel & Conference Center.

www.ohdsi.org/ohdsi2022symposium
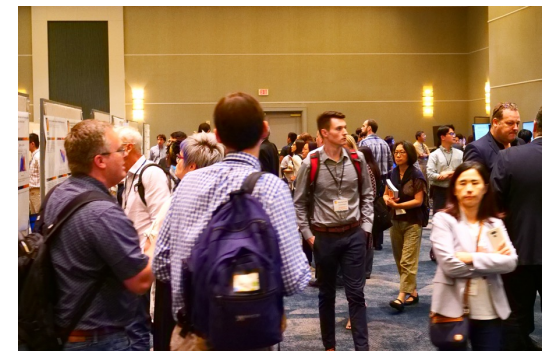
Workgroup Activities
Saturday, Oct. 15, and Sunday, Oct. 16

Saturday, Oct 15

Start Time (ET)	End Time (ET)				
800	900	Tutorial	HADES Hack-a-thon: Part 1	Oncology WG	FHIR-OMOP: Terminologies Subgroup, Part 1
900	1000				
1000	1100				FHIR-OMOP: Increasing the Value of Data Through a Rich Set of Attributes
1100	1200				
1200	1300		Lunch	Lunch	Lunch
1300	1400		Methods Research (PLE/PLP)	Oncology WG (continued)	FHIR-OMOP: Data Model Harmonization Subgroup
1400	1500				
1500	1600			Natural Language Processing	FHIR-OMOP: Oncology Subgroup
1600	1700				
1700	1800				FHIR-OMOP: Terminologies Subgroup, Part 2
1800	1900				

Sunday, Oct 16

Start Time (ET)	End Time (ET)				
800	900	All-Hands Workgroup Meeting			
900	1000				
1000	1100				
1100	1200				
1200	1300	Lunch		Lunch	Lunch
1300	1400	Phenotype Evaluation	HADES Hack-a-thon: Part 2	Education	CDM and Data Quality
1400	1500				
1500	1600			Health Equity	
1600	1700				

Hotel Block Rooms Available

Rooms Available for Oct. 13 and Oct. 14
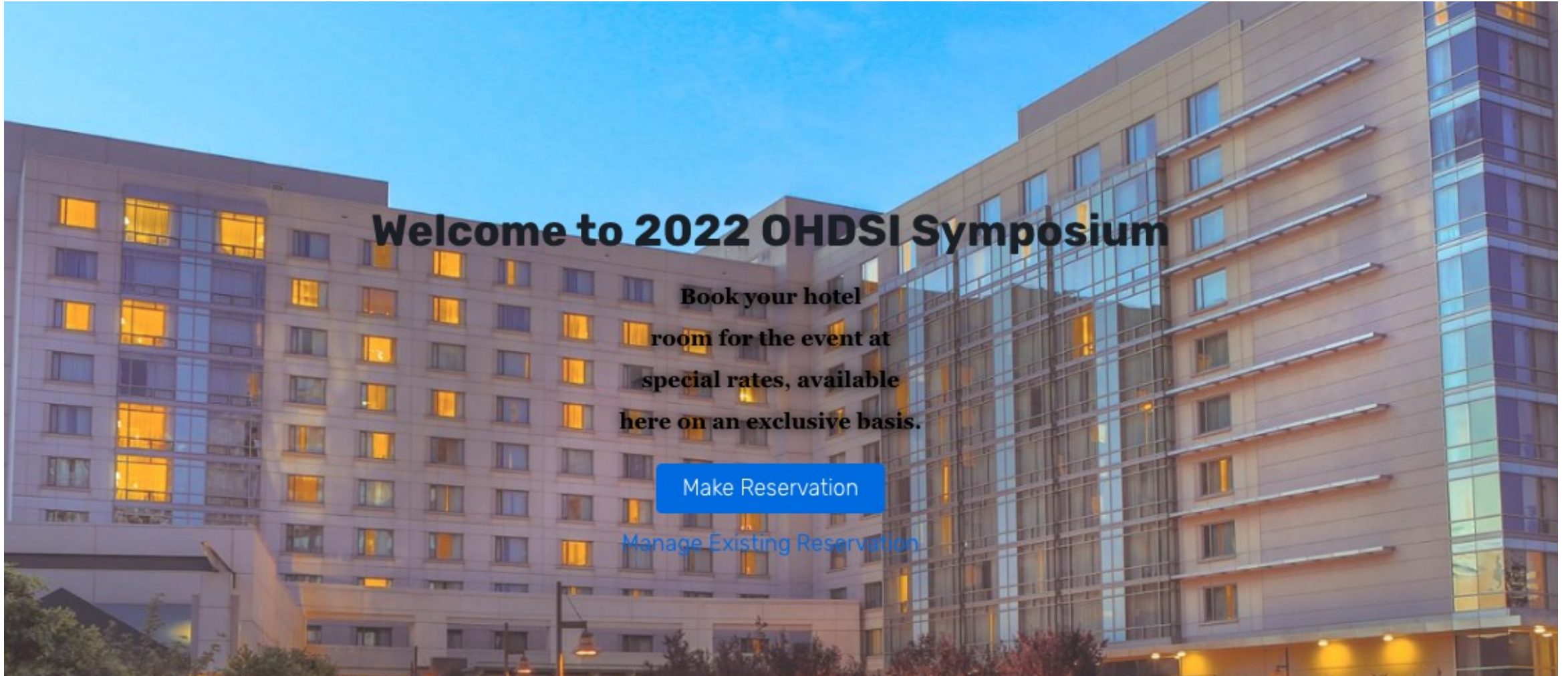
2022 OHDSI Symposium

OHDSI Community Calls ⌄ **Events & Past Collaborations** ⌄ **Learn About & Join OHDSI Workgroups** ⌄ **This Week In OHDSI** **EHDEN Academy** ⌄

Annual Report: Our Journey **Publications** **Support & Sponsorship** ⌄ **OHDSI2022 Symposium** ⌄ **Newsletters** ⌄ **Follow OHDSI on Social** ⌄

2022 OHDSI Symposium

Oct. 14-16 · Bethesda North Marriott Hotel & Conference Center

We are thrilled to announce that registration for the 2022 OHDSI Symposium, which will be held Oct. 14-16 at the Bethesda North Marriott Hotel & Conference Center, is now open!

It is so exciting to bring our community back together this fall. Our collaborator showcase will return; please click the link to see how you can take part in our poster presentations, software demos and lightning talks. The full agenda for our conference is still being developed, so please continue to check the OHDSI website (www.ohdsi.org) and our social platforms for updates as we plan for the 2022 Symposium.

The main conference will be held Friday, Oct. 14. A full-day tutorial will be held Saturday, Oct. 15, while other community activities will be held both Oct. 15 and Oct. 16.

Symposium Registration Details

Friday, Oct. 14 – Main Conference

Registration Fee: $500*
this is an open and inclusive event; if the registration fee represents a burden to you, please contact symposium@ohdsi.org.

Should you need to make changes or cancel your registration ticket, please follow the instructions you will receive on your Eventbrite confirmation upon registration completion. Please note that tickets can be refunded up until 7 days prior to the event; Eventbrite fees are not refundable.

Register For The Main Conference · Friday, Oct. 14

Saturday, Oct. 15 – Full-Day Tutorial: An Introductory Journey From Data To Evidence

Registration Fee: $300*
this is an open and inclusive event; if the registration fee represents a burden to you, please contact symposium@ohdsi.org.

Should you need to make changes or cancel your registration ticket, please follow the instructions you will receive on your Eventbrite confirmation upon registration completion. Please note that tickets can be refunded up until 7 days prior to the event; Eventbrite fees are not refundable.

Register For The Full-Day Tutorial · Saturday, Oct. 15

What Will Be Taught At This Tutorial?

Saturday, Oct. 15 and Sunday, Oct. 16 — Community Activities

A highlight of the OHDSI Symposium will be a full weekend of workgroup activities and meetings within the Bethesda North Marriott Hotel & Conference Center. You are now able to register for any workgroup sessions as long as there is no overlap between any two sessions; registration is free, but please do so early as this will be first-come, first-served due to room capacity.

See The Schedule & Agenda For Workgroup Activities · Weekend of Oct. 15-16

Register For Workgroup Activities · Weekend of Oct. 15-16

Hotel Information and Sleeping Room Block

Hotel: Bethesda North Marriott Hotel & Conference Center
Address: 5701 Marinelli Road, Rockville, Maryland, 20852
Hotel Main Number: (301) 822-9200
Reservations Toll Free: (877) 212-5752
Reservations Local Phone: (301) 822-9200

This year, OHDSI is holding a sleeping room block for the nights of Oct. 13 and 14 with a special room rate of $179 plus taxes. Please note that all sleeping rooms are on a first-come, first-served basis. To help us in the planning process, we ask that you do not cancel your hotel room ordered through the OHDSI Room Block. If you must cancel, please let us know prior to Thursday, Sept. 1, so that we can offer the room to others who may need one. Once the room block is full, or if specific nights are sold out, you may make additional room reservations on the hotel's website or by calling the hotel phone number above. Please note that OHDSI is not holding any sleeping rooms on Saturday, Oct. 15. Therefore, please call the hotel phone number or make this reservation online should you plan to stay Saturday night.

ohdsi.org/ohdsi2022symposium

Mapping of complex constructs in OMOP CDM

PRESENTER: Alexander Davydov

BACKGROUND

There is a growing need in the OHDSI community for the conversion of new types of data, different from the typical electronic health record systems or administrative claims data. This type of data is typically organized as entity-attribute-value (EAV) records, where the entity is either a question or a variable, the attribute is the link, and the value or answer is the value (table 1).

The ETL to convert them to OMOP CDM records is more complicated and requires specific solutions. In EAV, records often come in the form of variable/value or question/answer pairs. In the OMOP CDM, such data can be handled in the MEASUREMENT and OBSERVATION tables. The variable/question becomes the main concept, and the value/answer may be a value concept or not a concept at all.

Currently, there exist several approaches to relate different records in OMOP CDM: (i) through the SOURCE_TO_CONCEPT_MAP table; (ii) through the CONCEPT table in combination with the CONCEPT_RELATIONSHIP table and dedicated relationship IDs ("Maps to", "Maps to value", "Maps to unit", etc.). Both of these methods have certain limitations (table 2) that become relevant while converting EAV-modeled data.

METHODS AND RESULTS

In order to support and harmonize such data conversions, model changes in the structural organization of vocabulary mapping are required. It is crucial that mapping incorporates multiple source concepts, multiple target concepts of different domains, and data types (numeric, date, string). Here, we present two new solutions in this respect.

WIDE MAPPING table

The format is different in the way that mappings to multiple target entities will be performed by adding all the respective fields needed rather than creating the multiple rows and leveraging through the relationship_id/source_vocabulary_id (table 3).

Despite the promises, there are still some disadvantages and open questions:

- The table has a machine-readable format. Any attempts to add the descriptions will result in an excessive number of fields placed in one row so that users may have difficulties while building or looking up mappings.
- Source_concept_id key of the event table may be replaced by a foreign key back to the row in the wide mapping table, but then the table would become a reference for the source data. Currently, only the concept table plays this role.
- It's not clear how to represent text strings as a part of the source data.
- It forces users to create custom 2bil+ concepts out of the source data.
- Usually, the units of measure are separately coded in an additional field. The addition of the source_unit field to the wide mapping table gets us to a combinatorial explosion in most of the real-world data sources, even though it might be useful for controlled vocabularies and clean sources.
- The concept of the wide mapping table is to provide ETL with machine-readable instructions on how and where to extract the numeric value from. Additionally, it helps to differentiate the cases when there is no need to extract them (NULL numeric field). The ETL logic around it may be even more complex than the one that is currently used.

Remember Type Concept consolidation in 2020?

Concept Type consolidation

Stable ETLs

This time we really need **your** input.

Alexander Davydov, MD
Christian Reich, MD, PhD

ODYSSEUS DATA SERVICES   OHDSI

NUMERICAL RELATIONSHIP groups

Is an alternative solution that could be introduced into the concept_relationship table that indicate which attributes (value, unit, status, operator, modifier) belong to which "Maps to" relationship. Using the groups, the ETL scripts will automatically sort out one-to-many mappings and respective relationships into the target event records. In order to support the mappings of pre-coordinated pairs, the synthetic merged entity-value pairs still should be created. The mapping process to a field other than event_concept_id, will be leveraged using the new relationship_ids ("Maps to status", "Maps to operator", "Maps to modifier"). The most valuable benefit of this approach is that the mappings are still organized in many rows rather than in a single row and may be easily looked up by users. Also ETL logic and OMOP structure will not be affected that much. However, there are some points to be addressed still: mapping of (and to) ranges, numerics, strings and dates. The representation of these mappings probably still requires introduction of additional fields.

Type	Variable / Question	Value / Answer
Lab tests with the qualitative result	SARS-CoV-2 (COVID-19) IgA+IgM [Presence] in Serum or Plasma by Immunoassay	Equivocal / Negative / Positive
Historic facts	Family history of clinical finding	Myocardial infarction
Cancer stages and assessment measures	FIGO Stage (2018 FIGO Cancer Report)	I: Tumor confined to ovaries or fallopian tube(s)
	Circumferential Resection Margin (CRM)	100 mm or greater
Survey instruments created for specific projects (UK Biobank, All Of US PPI)	Has a doctor told you that you have any of the following problems with your eyes?	Macular degeneration
	How often did you use cannabis?	1-5 times per week
Surveys by itself (PhenX, PROMIS)	Because of your problem, do you feel frustrated	No / Sometimes / Yes
	Smoking helps me concentrate	Not at all / Somewhat / Very much

Table 1. Examples of EAV-structured data

Use case	Example	Issue
One-to-many "splitting" mappings through multiple relationships	"Maps to" and "Maps to value" pairs: "History of" + value of "COVID-19 vaccine" together with "SARS-COV2 PCR test" + value of "POS"	It is ambiguous which "Maps to" belongs to which "Maps to value", and the standard ETL process will inflate the records
Many-to-one "merging" or "pre-coordination" mapping	HHV-6B seropositivity for Human Herpesvirus-6: False	Only a single code can be an input for a map. As a result, the ETL needs to apply a workaround and first merge the entity/value codes to map them to the target concept
	EuroQol five dimension three level self-care score: 3 (I am unable to wash or dress myself)	
Separate mapping of entities and values, which is possible only if the values are entity agnostic	Generic "Yes", "No" answers to questions; drugs, conditions and other self-sufficient concepts	Now this is managed by splitting the source codes into separate synthetic source vocabularies
Mapping to numeric content or numeric with a unit	CS Tumor Size of 32 mm	Currently, ETL needs to extract the numeric values and units from the text
Mapping of a range	Blood alcohol level of 100-119 mg/100 ml	Ranges are currently not supported
Mapping to a string	White sliced bread eaten	Currently, ETL needs to extract the values from the text
Mapping to a date	Birthdate of a relative: "1988-Sep-17"	Currently, ETL needs to extract the dates from the source

Table 2. Limitations of existing system

Table 3. The structure of the wide mapping table and mapping examples

Source				Target								
Source Concept	Question/Variable	Answer/Value	Range	Standard Concept	Numeric	Operator	Error	Unit Concept	Value Concept	String	Condition Status Concept	Visit Concept
Ambulatory procedures - lithotripsy				Lithotripsy								Ambulatory Surgical Center
	CS Tumor Size	\d	001 - 988 millimeters (mm) (Code exact size in mm)	Estimated Tumor Size	\d			millimeter				
Documentation of patients with primary headache diagnosis and imaging other than ct or mri obtained				Headache							Primary diagnosis	
				Imaging								
Evidence of alcohol involvement determined by blood alcohol level of 100-119 mg/100 ml				Ethanol [Mass/volume] in Blood	110	10		milligram per deciliter				
Home visit, phototherapy services (e.g., bili-lite), including equipment rental, nursing services, blood draw, supplies, and other services, per diem				Home visit, phototherapy services (e.g., bili-lite), including equipment rental, nursing services, blood draw, supplies, and other services, per diem								Home Visit
	Wears glasses or contact lenses	Yes		Abnormal vision								
				Uses visual aid								
	Age started wearing glasses or contact lenses	\d (e.g. 15)		History of event **longer than 10 years ago**					Uses visual aid			
	Type of sliced bread eaten	white		Food eaten						"white sliced bread"		

#OHDSISocialShowcase This Week

Implementing the OHDSI Community Approach to Phenotype a Complex Medical Condition in European Primary Care Data

Authors: Kristin Kostka[1,2], Evan Minty[3], Antonella Delmestri[1], Barrack Omondi[1], Martí Català[1], Edward Burn[1,5], Daniel Prieto-Alhambra[1,4], Annika M. Jödicke[1]

Affiliations: [1]Pharmaco- and Device Epidemiology, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, OX3 7LD, UK; [2]The OHDSI Center at the Roux Institute, Northeastern University, Portland, ME, US; [3]O"Brien Centre for Population Health, Faculty of Medicine, University of Calgary, CA; [4]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, NL; [5]Fundació Institut Universitari per a la recerca a l'Atenció Primària, Barcelona, ES

Introduction

Background: "Post-acute COVID-19 syndrome" or "long COVID" are persistent symptoms that continue for weeks or months following the acute COVID-19 disease. As the COVID-19 pandemic continues, long COVID poses a significant public health issue with potential to inflict mass disability [1]. Clinicians have varying familiarity in the characteristic symptoms associated with long COVID, creating challenges in defining and measuring this issue at scale.

The Atlantic
Long COVID Could Be a 'Mass Deterioration Event'
A tidal wave of chronic illness could leave millions of people incrementally worse off.

The Guardian
Two million people in UK living with long Covid, find studies

POLITICO
Can long Covid lead to death? A new analysis suggests it could

Figure 1. A sampling of media coverage of long COVID from Spring 2022

Objective: To follow OHDSI best practices for developing a long COVID phenotype and apply them to UK OMOP CDM-mapped primary care data.

Methods

Engaging the Community: We partnered with the OHDSI Phenotype Development & Evaluation Workgroup to run a Long COVID phenotyping hackathon on December 7, 2021. In the hackathon, we used the World Health Organization (WHO) Delphi consensus of the clinical case definition of post COVID-19 condition [2]. We assembled concept sets for the 25 individual symptoms using a consistent process (Figure 2). Each concept set expression was inspected through use of PHOEBE[4], PheValuator[5], and available literature.

Clinical definition → Initial Concept Set Expression (via ATHENA and existing studies) → Review with PHOEBE and/or PheValuator → Review with Clinicians → Create Cohort Definitions → Generate Cohort Diagnostics
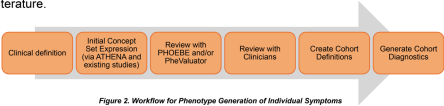
Figure 2. Workflow for Phenotype Generation of Individual Symptoms

Tapping into Cohort Diagnostics: We ran an initial CohortDiagnostics[3] package on a large database of UK primary care electronic health records, Clinical Practice Research Datalink (CPRD) AURUM mapped to OMOP CDM V5.3. The study period started on 1 January 2020 and ended at the last available date (11 Mar 2021). We then used the symptom code lists to iteratively constructed cohort definition parameters to generate 125 cohorts. To enter any cohort, persons were required to be over 18 years of age, have a qualifying COVID diagnosis or positive PCR test and at least 180 days of prior observation time. *(Note: Acute COVID entry criteria were reused from prior Oxford research by Burn et al.)* Additional inclusion criteria consisted of no history of the specific symptom prior to index (- 90 days, -180 days) and a time window of symptom persistence (+28 days, +90 days after diagnosis or test). In a subset of symptoms, we explored the use of a run-in time window (-7 days, -14 days) where symptoms may present prior to clinical confirmation of acute COVID-19. After initial clinical review of the individual phenotypes, a composite long COVID phenotype was assembled.

Results

- The 1-day community hackathon produced:
  - 7 final clinical symptom concept set expressions meeting the OHDSI best practices
  - 9 drafted clinical symptom concept set expression for further review with OHDSI diagnostics
  - 9 clinical symptom concept set expressions to be developed.
  - (… and Gowtham broke PHOEBE. ☺)
- 1 WHO symptom (post-exertional malaise/fatigue) was dropped from the concept set process due to insufficient use of concepts in primary care data.
- The 18 concept sets were later run through PHOEBE and reviewed by clinical input. Iterative results are stored in the OHDSI Phenotype WG Long COVID channel on OHDSI MS Teams.
- In the CohortDiagnostics review, a total of 458,975 persons with COVID-19 diagnosis or a positive test met the cohort entry criteria (C124).
- The most common persistent symptoms included shortness of breath (n=4005; C45), anxiety (n=3378; C6), joint pain (n=3340; C14), cough (n=3275; C32), abdominal pain (n=2651; C1) and depression (n=2552; C10).
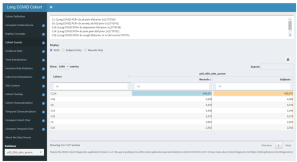- Cohort counts were impacted by prior history, symptom persistence, and run-in windows.

Figure 3. CohortDiagnostics Shiny Application for Study

SCAN ME
Let's collaborate!
Visit our study repo

Conclusion

The OHDSI community approach to phenotyping provides a robust framework to evaluating a complex medical condition, such as long COVID. We observed differences in cohorts based on logic changes in prior follow-up time, time for symptom persistence, gender, and age. Our findings can help researchers understand the impact of fluctuating clinical logic on describing and measuring long COVID at scale.

Reference: [1] U.S. Government Accountability Office. Science & Tech Spotlight: Long COVID. [cited 16 Jun 2022]. Available: https://www.gao.gov/products/gao-22-105666. [2] Soriano JB, Murthy S, Marshall JC, Relan P, Diaz JV, WHO Clinical Case Definition Working Group on Post-COVID-19 Condition. A clinical case definition of post-COVID-19 condition by a Delphi consensus. Lancet Infect Dis. 2022;22: e102-e107. doi:10.1016/S1473-3099(21)00703-9. [3] Rao G, Schuemie M, Ryan P, Weaver J, Gilbert J. CohortDiagnostics: Diagnostics for OHDSI Cohorts. 2022 [cited 5 May 2022]. Available: https://ohdsi.github.io/CohortDiagnostics/

Contact: kristin.kostka@ndorms.ox.ac.uk

TUESDAY

Implementing the OHDSI Community Approach to Phenotype a Complex Medical Condition in European Primary Care Data
Lead: Kristin Kostka

@OHDSI   www.ohdsi.org   #JoinTheJourney   ohdsi

#OHDSISocialShowcase This Week

The use of data-driven vs. clinical based propensity score in COVID-19 vaccine safety research:

Association between thrombosis with thrombocytopenia syndrome (TTS) or thromboembolic events (TE), and COVID-19 vaccines

PRESENTER: **Xintong** Li

INTRO

- Propensity score (PS) have been widely used in observational studies to reduce confounding by indication
- Clinical knowledge based vs. data-driven PS

METHODS

Data source: OMOPed data from 5 European counties: France, Germany, Netherlands, Spain, and the United Kingdom) and two from the United States.

Cohort study:
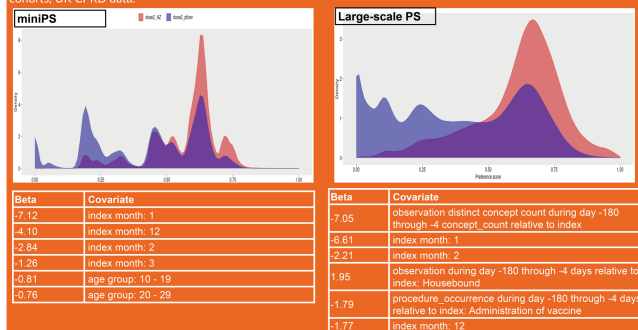
- Target: adenovirus-based
- Comparator: mRNA

Analysis:

- miniPS: clinically-driven
- Large-scale PS: data-driven, L1 regularized logistic regression
- 1-to-4 matching

Diagnostics:

1. Measured confounding: Covariate balance after propensity score matching (SMD < 0.1)
2. Power: minimal detectable relative risk in the matched cohorts
3. Systematic error: using negative control outcomes

Figure 1. Propensity score distribution covariates with top 6 absolute values of Beta, 2nd dose Vaxzevria and Comirnaty cohorts, UK CPRD data.

miniPS — dose1_AZ dose2_pfizer

Large-scale PS

Beta	Covariate
-7.12	index month: 1
-4.10	index month: 12
-2.84	index month: 2
-1.26	index month: 3
-0.81	age group: 10 - 19
-0.76	age group: 20 - 29

Beta	Covariate
-7.05	observation distinct concept count during day -180 through -4 concept_count relative to index
-6.61	index month: 1
-2.21	index month: 2
1.95	observation during day -180 through -4 days relative to index: Housebound
-1.79	procedure_occurrence during day -180 through -4 days relative to index: Administration of vaccine
-1.77	index month: 12

While selected confounders were balanced in clinical-based PS after matching, other potentially relevant covariates remained unbalanced, suggesting residual confounding

Figure 2. Before and after matching SMD, 2nd dose Vaxzevria and Comirnaty cohorts, UK CPRD data.

miniPS: 73 variables included

SMD=0.418: observation during day -180 through -4 days relative to index: Informed consent given for treatment

SMD=0.174: observation during day -180 through -4 days relative to index: Housebound

Large-scale PS: 5,586 variables included
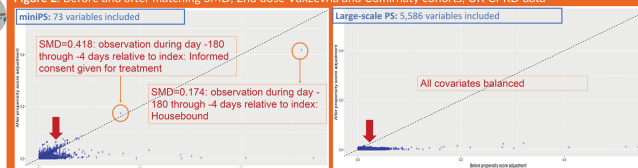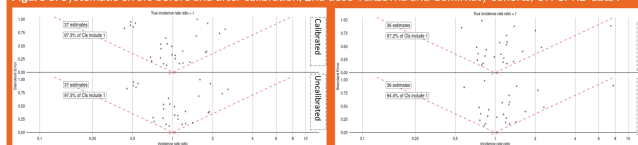
All covariates balanced

Figure 3. Systematic errors before and after calibration, 2nd dose Vaxzevria and Comirnaty cohorts, UK CPRD data.

RESULTS

- PS distribution
- Before and after matching SMD
- Systematic error using negative control outcomes

CONCLUSIONS

- Index month and age have high impact for both clinical based and data-driven propensity scores.
- Clinical-based PS: balanced on selected variables, but not other covariates
- Large-scale PS: all covariates were well-balanced after matching
- Performance on controlling systematic errors were similar
- Cons of large-scale: computing time (30mins vs. 6 hours on a 250,000 down sampling cohort)

Table 1. Summary of the covariate balance for both propensity scores.

Database	Target	Comparator	No SMD > 0.1 after matching Mini PS	No SMD > 0.1 after matching Large-scale PS
UK CPRD Aurum	Vaxzevria 1st	Comirnaty 1st	✕	✓
UK CPRD Aurum	Vaxzevria 2nd	Comirnaty 2nd	✕	✓
Germany DA	Janssen	Comirnaty 1st	✕	✓
NL IPCI	Vaxzevria 1st	Comirnaty 1st	✕	✓
US OpenClaims	Janssen	Comirnaty 1st	✓	✓
US OpenClaims	Janssen	Spikevax 1st	✓	✓

*CPRD AURUM: Clinical Practice Research Datalink (CPRD) Aurum, United Kindom; IPCI: Integrated Primary Care Information (IPCI), The Netherlands; DA Germany: IQVIA Disease Analyser (DA) Germany; US OpenClaims: Medical and Institutional Claims (Dx and Hx); SMD: standardized mean difference.

Xintong Li[1], Edward Burn[1,2], Prieto-Alhambra[1,3]

1.Centre for Statistics in Medicine (CSM), Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDROMS), University of Oxford, UK
2.Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain
3. Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute, Faculty of Health and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain

NDORMS OXFORD OHDSI

WEDNESDAY

The use of data-driven vs. clinical based propensity score in covid-19 vaccine safety research
Lead: Xintong Li

#OHDSISocialShowcase This Week

Norwegian registry data on to OMOP CDM - mapping challenges and opportunities for pregnancy studies

PRESENTER: **Eimir Hurley**

INTRODUCTION:
In October 2021, University of Oslo was granted funds from the EHDEN - Data Partner Call to map five main Norwegian health registries onto the OMOP CDM (Figure 1).

Our aim is to enable collaboration with other data providers using this CDM, especially for perinatal pharmaco-epidemiological studies.

METHODS
The following Norwegian registries data (2018-2020) were mapped to the OMOP CDM v5.3.1:

- Medical Birth Registry of Norway (MBRN)
- Norwegian Prescription Database (NorPD)
- Norwegian Patient Registry (NPR)
- Norwegian Surveillance System for Communicable Diseases (MSIS)
- Norwegian Immunisation Registry (SYSVAK)

Overview of our mapping procedure is presented in Figure 2.

We enrich pregnancy data in OMOP format

OHDSI tools were effectively utilized

No standard concepts for important pregnancy related variables

Pregnancy extension table for future version of the OMOP CDM?

Figure 1: Five Norwegian registries mapped onto OMOP at the University of Oslo, Norway

5.4 million inhabitants

Medical Birth Registry of Norway (MBRN)
Norwegian Prescription Database (NorPD)
Norwegian Patient Registry (NPR)
Norwegian Immunisation Registry (SYSVAK)
Norwegian Surveillance System for Communicable Diseases (MSIS)

Unique personal identifiers facilitating linkage between registries
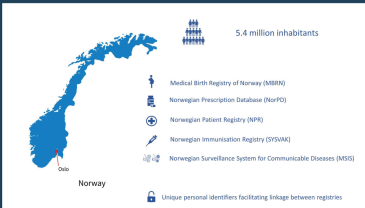
Figure 2: Overview of mapping procedure
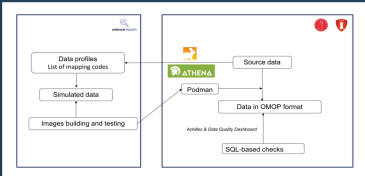
Figure 3: Data quality assessment on MBRN

DATA QUALITY ASSESSMENT

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	2013	0	2013	100%	287	0	287	100%	2300	0	2300	100%
Conformance	674	5	679	99%	104	0	104	100%	778	5	783	99%
Completeness	384	2	386	99%	9	6	15	60%	393	8	401	98%
Total	3071	7	3078	100%	400	6	406	99%	3471	13	3484	**100%**

List of OMOP tables:

- **Person**
- **Fact_relationship**
- **Death**
- **Provider**
- **Visit_occurrence**
- **Condition_occurrence**
- **Procedure_occurrence**
- **Drug_exposure**
- **Measurement**
- **Observation**
- **Observation_period**

RESULTS

5 508 030 individuals (2018-2020)

- 720 765 pregnancies
- 452 831 mothers
- 440 731 fathers
- 695 569 children

237 non-standard codes were mapped to standard concepts:

- 67 pregnancy related codes
- 48 speciality related codes
- 48 communicable disease related codes
- 46 vaccine related codes
- 17 drug related codes
- 3 procedure related codes

40 custom concepts were introduced to accommodate terminologies that were not supported by OMOP vocabularies. 36/40 codes provide vital information about pregnancies:

- Previous miscarriages before 12 weeks of gestations
- Early preeclampsia
- Sometimes smoking before pregnancy
- Daily smoking at the start of pregnancy
- Hemorrage more than 1500ml during delivery

Study team:
Nhung Trinh, Jared Houghtaling, Fabian LM Bernal, Eimir Hurley, Emma Gesquiere, Lars Halvorsen, Hedvig ME Nordeng

---

THURSDAY

Norwegian registries onto OMOP Common Data Model: mapping challenges and opportunities for pregnancy studies
Lead: Elmir Hurley

#OHDSISocialShowcase This Week

De-identification of Clinical Notes for Patients with Infectious Diseases and Topic Modeling using Latent Dirichlet Allocation

PRESENTER: **Junhyuk** Chang

INTRO
- Infectious disease-related information is usually recorded in the form of free-text, which needs natural language processing (NLP) to apply.
- However, most of free-text is containing protected health information (PHI) that should be de-identified.
- In this study, we applied the NLP to confirm the distribution of infection-related information after de-identifying PHI in admission notes.

METHODS
1. Data preparation
- Ajou University Medical Centre database
- Inclusion criteria
1) Admitted from Jan 2012 - Dec 2021.
2) Diagnosed with infectious disease within ±2 days from the admission date.
   - Infectious disease diagnosis
   : SNOMED code '40733004 (Disorder due to infectious disease)' and its sub-hierarchy codes

2. PHI identification and de-identification
- We compared 1,000 admission notes that were randomly selected with the HIPAA PHI list to identify the potential PHI entity.
- Two approaches to de-identify PHI entities

1) Dictionary-based approach	2) Rule-based approach
• For name, country, and hospital entities	• For other PHI patterns

3. Feature identification using topic modeling
- Tokenization
  - By unigram
  - Descriptive analyses for frequency
- Latent Dirichlet allocation (LDA)
  - Describing documents by clustering words based on the frequency
  - Perplexity score to decide an optimal n of topics

RESULTS
Extract admission notes and PHI de-identification
- We extracted patients and their admission notes.
- We identified PHI entities and their patterns.

44,415 patients	61,379 Admission notes	9 PHI 21 Pattern

- Constructed dictionaries (dictionary : cases)
  - Name : 47,696, Country : 241, Hospital : 45,932 (regular expression rules to de-identify showed in the abstract).

Infectious disease can be screened and detected through natural language processing after de-identifying patient health information
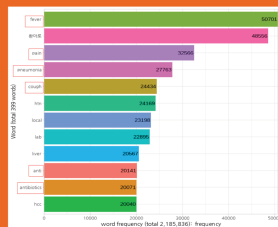

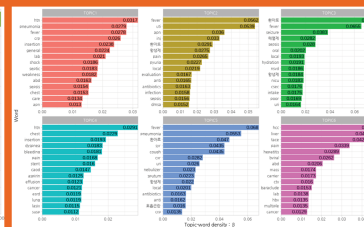**Figure 1.** Word frequency plot for total documents

Figure 2. Word density plot for four topics.

Scan QR to **download** the **abstract** or **poster.**

Descriptive summary
- "**fever**" has the highest frequency (50,701/2,185,836 ; 2.3%) (Figure 3).
- **Infectious disease related words** (red box) also showed high frequency.

LDA topic modeling
- Decided optimal topic number
  - 5–9 topics were the optimal topic number according to the perplexity score
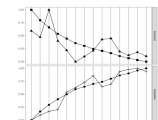 - **6 topics** for a clear explanation of semantic meanings

Figure 3. Perplexity scores plot

- Figure 2 shows the most frequently identified words per each topic.
- Clustered word per each topic related below.

Topic 1 Sepsis	Topic 2 Urinary tract infection	Topic 3 Pediatric infection
Topic 4 Surgical infection	**Topic 5** Respiratory infection	**Topic 6** Viral infection

- Relevance of clustered words per each topic (Figure 4).

Figure 4. Topic distance map and relevant terms for the topic 2

CONCLUSION
- In this study, we extracted sign and symptoms related to infectious disease from deidentified clinical records using natural language processing technique.
- This framework can be used for future research such as data standardization of infectious disease and cohort phenotyping.

Junhyuk Chang[1], Jimyung Park[1], Chungsoo Kim[1], Rae Woong Park[1,2]

[1]Department of Biomedical Sciences, Ajou University Graduate School of Medicine
[2]Department of Biomedical Informatics, Ajou University School of Medicine

---

FRIDAY

PHAROS, Platform for Harmonizing and Accessing Data in Real-time on Infectious Disease Surveillance Based on OMOP-CDM in Korea
Lead: Chungsoo Kim

@OHDSI   www.ohdsi.org   #JoinTheJourney   ohdsi

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

the journey