Week 2
Workgroup 2023 OKRs and Phenotype Phebruary Updates

OHDSI Community Call
Feb. 14, 2023 • 11 am ET

Upcoming OHDSI Community Calls

Date	Topic
Feb. 21	Phenotype Phebruary Weekly Update + Workgroup Plans for 2023
Feb. 28	Phenotype Phebruary Weekly Update + Workgroup Plans for 2023
Mar. 7	Save Our Sisyphus (SOS) Research Idea Presentations
Mar. 14	OHDSI Debates
Mar. 21	Recent Publications
Mar. 28	SOS Week 1 Tutorial: Initiating A Network Study

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the **EHDEN Academy team** on the recent announcement that Academy users are spread out **over 100 different countries**.

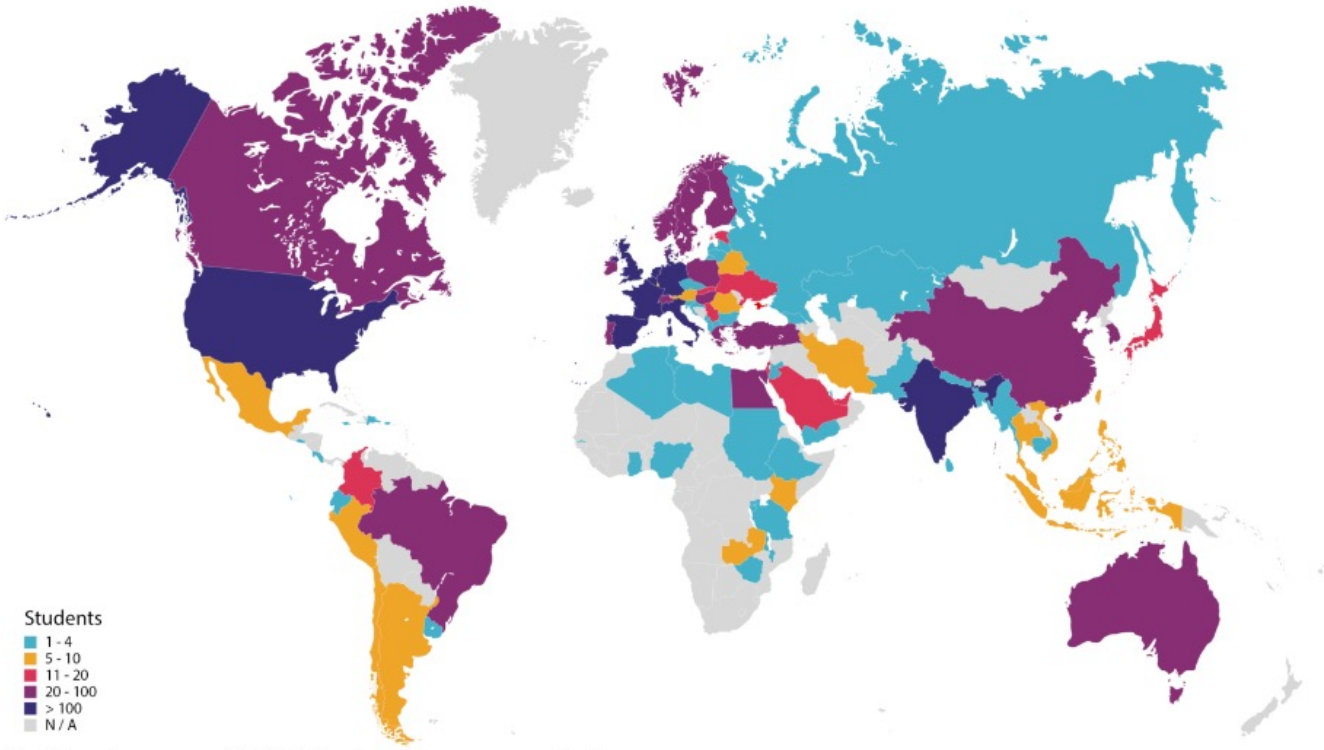


Fig 1. Current coverage of EHDEN Academy course users worldwide

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!

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Wednesday	11 am	Open Source Community
Wednesday	12 pm	Health Equity Journal Club
Thursday	12 pm	HADES
Thursday	1 pm	OMOP CDM Oncology Vocabulary/Development Subgroup
Thursday	7 pm	Dentistry
Friday	9 am	GIS – Geographic Information System Development
Friday	1 pm	Clinical Trials
Monday	10 am	Africa Chapter

ohdsi.org/workgroups

European Symposium: July 1-3, 2023

OHDSI HADES releases: OhdsiShinyModules v1.0.2

OhdsiShinyModules 1.0.2 | Reference | Contribute | Articles | Changelog | hadesLogo

OhdsiShinyModules

R-CMD-check passing | codecov 83%

OhdsiShinyModules is part of HADES.

OhdsiShinyModules is an R package containing shiny modules that can be used within shiny result interfaces.

The OHDSI tools often provide shiny interfaces for viewing and exploring results. Many of these shiny apps have overlapping features. To ensure consistency we have created a repository containing useful shiny modules that can be used in multiple result explorers.

Current Modules

- about module: this contains information about the shiny viewer and the types of OHDSI analyses
- prediction module: a module for exploring patient-level prediction results that were developed usign the OHDSI PatientLevelPrediction package

Technology

OhdsiShinyModules is an R package that uses the R shiny library.

Links

Ask a question

License

Apache License 2.0

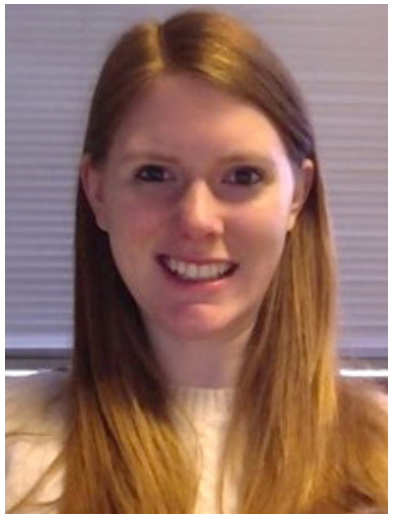
Citation

Citing OhdsiShinyModules

Developers

Jenna Reps

Maintainer



OHDSI HADES releases: ShinyAppBuilder v1.1.1

ShinyAppBuilder 1.1.1 | Reference | Articles | Changelog | HADES

ShinyAppBuilder

R-CMD-check passing | codecov 82%

ShinyAppBuilder is part of HADES.

Introduction

Create shiny apps using modules from OhdsiShinyModules or custom modules

Examples

To create a shiny viewer to explore CohortDiagnostic results, Characterization results, PatientLevelPrediction results and CohortMethod results:

```
# install dependencies
remotes::install_github('ohdsi/ResultModelManager')
remotes::install_github('ohdsi/ShinyAppBuilder')
```

Links

Browse source code

Report a bug

Ask a question

License

Apache License 2.0

Citation

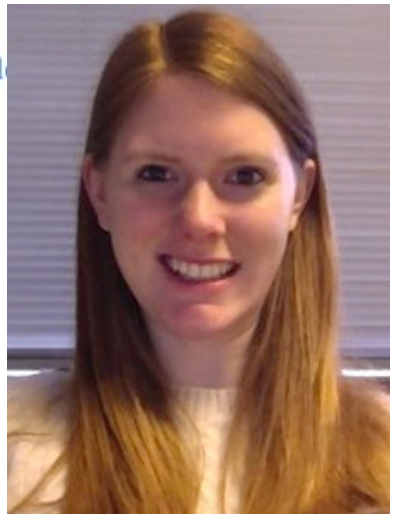
Citing ShinyAppBuil

Developers

Jenna Reps
Author, maintainer

Josh Ide
Author

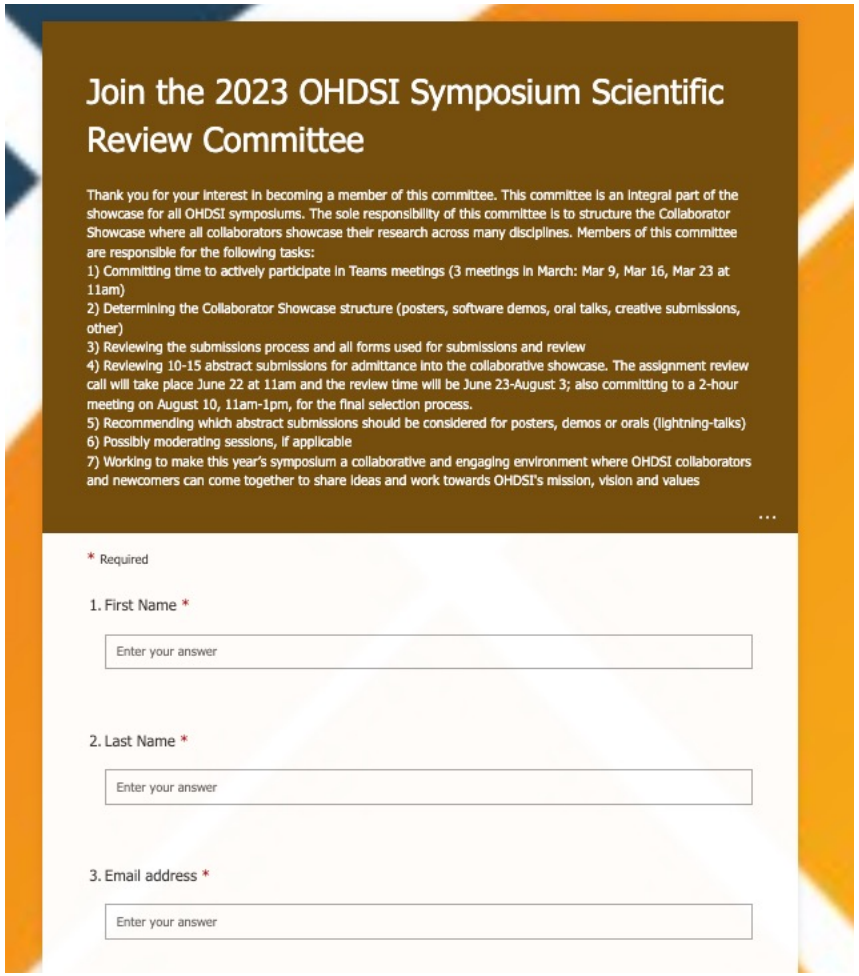
Jamie Gibert
Author



Join The #OHDSI2023 Scientific Review Committee

#OHDSI2023 is coming Oct. 20-22, and we are looking for collaborators to join the scientific review committee. **Elisse Katzman** has opened the signup form to join the committee, and the first meeting is scheduled for March 9.

Deadline is Feb. 28.



Join the 2023 OHDSI Symposium Scientific Review Committee

Thank you for your interest in becoming a member of this committee. This committee is an integral part of the showcase for all OHDSI symposiums. The sole responsibility of this committee is to structure the Collaborator Showcase where all collaborators showcase their research across many disciplines. Members of this committee are responsible for the following tasks:
1) Committing time to actively participate in Teams meetings (3 meetings in March: Mar 9, Mar 16, Mar 23 at 11am)
2) Determining the Collaborator Showcase structure (posters, software demos, oral talks, creative submissions, other)
3) Reviewing the submissions process and all forms used for submissions and review
4) Reviewing 10-15 abstract submissions for admittance into the collaborative showcase. The assignment review call will take place June 22 at 11am and the review time will be June 23-August 3; also committing to a 2-hour meeting on August 10, 11am-1pm, for the final selection process.
5) Recommending which abstract submissions should be considered for posters, demos or orals (lightning-talks)
6) Possibly moderating sessions, if applicable
7) Working to make this year's symposium a collaborative and engaging environment where OHDSI collaborators and newcomers can come together to share ideas and work towards OHDSI's mission, vision and values

* Required

1. First Name *

Enter your answer

2. Last Name *

Enter your answer

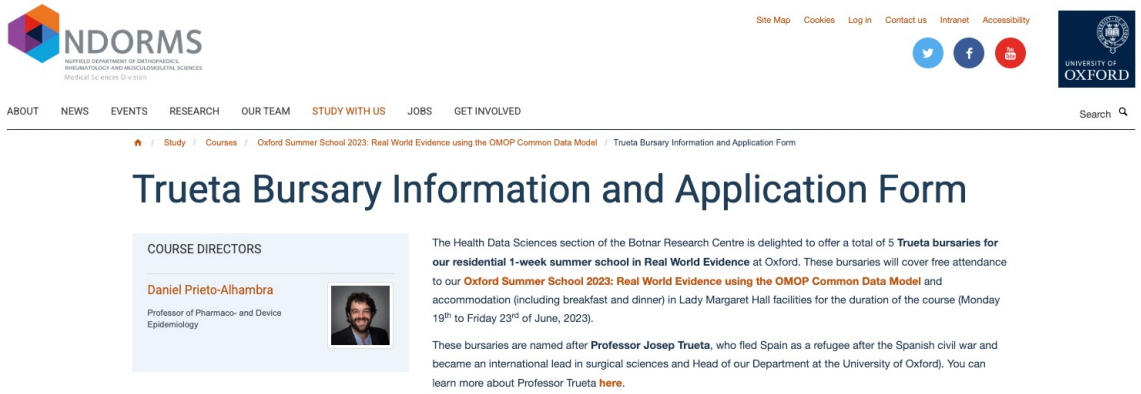
3. Email address *

Enter your answer

bit.ly/OHDSI2023ScientificReview

New Opportunity For Oxford RWE Summer School

Dani Prieto-Alhambra shared that the Health Data Sciences section of the Oxford University Botnar Research has created **five Trueta bursaries for the residential one-week summer school** in Real World Evidence, held June 19-23, 2023. These bursaries will cover free attendance to the Oxford Summer School 2023: Real World Evidence using the OMOP Common Data Model and accommodation (including breakfast and dinner) in Lady Margaret Hall facilities for the duration of the course.



NDORMS

Site Map Cookies Log in Contact us Intranet Accessibility

UNIVERSITY OF OXFORD

ABOUT NEWS EVENTS RESEARCH OUR TEAM STUDY WITH US JOBS GET INVOLVED Search

Study / Courses / Oxford Summer School 2023: Real World Evidence using the OMOP Common Data Model / Trueta Bursary Information and Application Form

Trueta Bursary Information and Application Form

COURSE DIRECTORS

Daniel Prieto-Alhambra

Professor of Pharmaco- and Device Epidemiology

The Health Data Sciences section of the Botnar Research Centre is delighted to offer a total of 5 **Trueta bursaries for our residential 1-week summer school in Real World Evidence** at Oxford. These bursaries will cover free attendance to our **Oxford Summer School 2023: Real World Evidence using the OMOP Common Data Model** and accommodation (including breakfast and dinner) in Lady Margaret Hall facilities for the duration of the course (Monday 19th to Friday 23rd of June, 2023).

These bursaries are named after **Professor Josep Trueta**, who fled Spain as a refugee after the Spanish civil war and became an international lead in surgical sciences and Head of our Department at the University of Oxford). You can learn more about Professor Trueta **here**.

To be eligible for these bursaries, you need to have completed previous training in a field relevant to **health data sciences***, and to fulfil at least one of the following criteria:

- **Be a refugee, stateless person, or asylum seeker**, or been otherwise forcibly displaced within or outside your region or country of origin
- **Residence in a low/middle-income country** as defined by the Organisation for Economic Co-operation and Development (OECD), listed here: **https://wellcome.org/grant-funding/guidance/low-and-middle-income-countries**
- **UK residents from a low income background**, as defined by the Department of Work and Pensions: **https://www.ethnicity-facts-figures.service.gov.uk/work-pay-and-benefits/pay-and-income/people-in-low-income-households/latest**
- **Current post-graduate (MSc or PhD) student**

We particularly encourage applications from people from ethnicities under-represented in the field of health data sciences, and from people with current or recent caring responsibilities.

2023 Health Data Science Black Internship Program

Dani Prieto-Alhambra announced that applications are open for the 2023 Health Data Science Black Internship Programme at the University of Oxford.

Closing date for registration is Feb. 27.

UNIVERSITY OF OXFORD

About the programme

The aim of our Black Internship Programme is to tackle the underrepresentation of Black people within the health data science sector. We are doing this by providing you with an opportunity to expand your knowledge around health data science and gain the experience you need to kickstart (or advance) your career in this field.

Our internships are a super way to gain hands-on experience, carrying out practical projects in the real world. It's a great way to find out about a rapidly advancing area of science, meet fellow interns, looks superb on a CV – and can open doors to new opportunities.

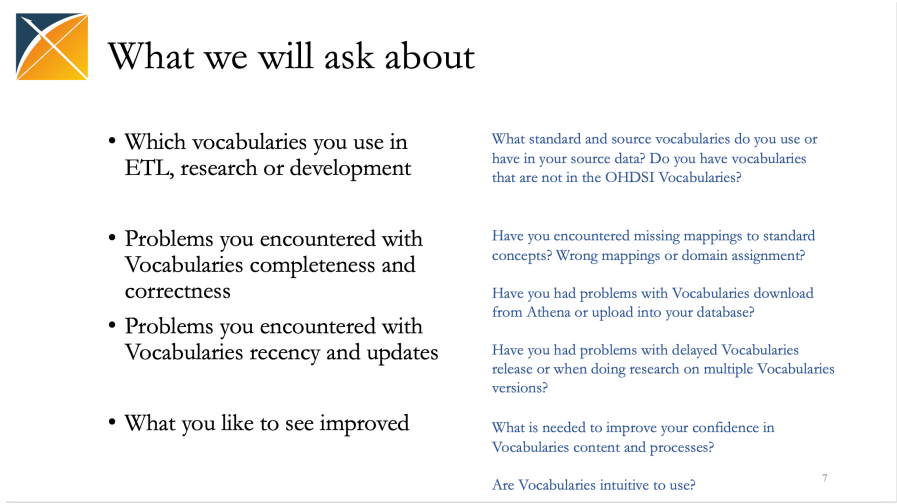
Planned for the summer of 2023, this internship programme will offer:

- 8-week paid internship
- Opportunity to join the wider 10,000 Black Interns community
- Certificate recognising intern achievements
- Opportunities across sectors in health data
- Customised learning pathway within HDR UK Futures
- Ongoing support post-programme from HDR UK Alumni Network and access to HDR UK Futures
- A real world data project developed by our host organisations
- Mentor and line manager
- Opportunity to shape next year's programme
- Cohort-building and training activities every Friday afternoon
- Team technical challenge and prize giving

Vocabulary Landscape Assessment

Anna Ostropolets introduced **a pair of vocabulary landscape assessment surveys** to directly inform which vocabularies and activities the vocabulary team prioritizes in 2023.

The deadline is Feb. 23.



Main vocab: bit.ly/3iTnyco || ETL/Data owners: bit.ly/3R7rYcm

Save Our Sisyphus Challenge

Save Our Sisyphus Challenge

The task of taking a research study from idea through design through execution through publication can seem a daunting challenge, much like rolling a boulder up a hill. That task is all the more challenging when researchers try to go it alone, as each step requires a distinct set of skills. Observational study design requires epidemiologic understanding and statistical methodological expertise. Implementing a study design requires statistical programming ability. Interpreting and reporting results requires domain knowledge of the clinical problem.

But when you are part of the OHDSI community, you never have to go it alone. And as a team effort, what seems an arduous task can become an efficient and effective process.

We are seeking important research questions that you want to contribute and participate in to take from idea to publication. The OHDSI community will provide support through every step of the process, working with you to design an appropriate protocol, implement a network analysis package, execute across OHDSI data partners, and prepare a manuscript for publication. Our goal is to collaboratively complete this network study over the course of 8 weeks across April and May, using the open-source tools and process that OHDSI has

https://forms.gle/DySfETJPtmwgquKv9

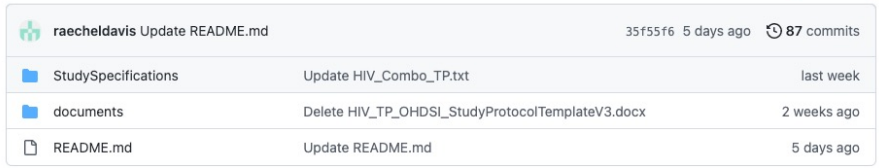
PhD Student Opportunity

Georgina Kennedy shared a recent opportunity for **a PhD student at the University of New South Wales, Sydney**, to join in a project to understand the current use and future potential of real-world data to measure, explain and respond to variation in clinical cancer care. This funded position comes with a living costs stipend, and a technical background is required.

georgina.kennedy@unsw.edu.au

Join This OHDSI Network Study

Rachael Davis is leading a network study **to characterize and evaluate trends in pathways for antiretroviral therapy (ART) for individuals who have been diagnosed with Human Immunodeficiency Virus (HIV) and treated persistently over two years**. She is seeking collaborators and data partners for this study.



README.md

Characterizing Patterns in Antiretroviral Therapy for Individuals with Human Immunodeficiency Virus (HIV)

Study Status | Repo Created

- Analytics use case(s): **Characterization**
- Study type: **Clinical Application**
- Tags: **OHDSI, HIV, Antiretroviral Therapy**
- Study lead: **Raechel Davis**
- Study lead forums tag: **redavis**
- Study start date: **January 2023**
- Study end date: **-**
- Protocol: **HIV_TP**
- Publications: **-**
- Results explorer: **-**

How to run

Requirements:

- A Database in the Common Data Model (>= Version 5)
- Configured ATLAS Environment with capability to execute CohortPathways and Characterization Tools

1. Import Study Specifications

Import the following specifications for each analysis into ATLAS by copying and pasting json .txt provided into the Utilities, Import Tab for each analysis:

- Cohort Pathway 1: **Single Ingredient Antiretrovirals**
- Cohort Pathway 2: **Combination Antiretroviral Medications**

note: the minimum cell count default is 1

- Characterization: **Characterization**

2. Execute Cohort Pathway 1, Cohort Pathway 2, and Characterization on CDM Transformed Database

Facilitate the analysis under the Execution tab and click 'Generate'

3. Return csv Results Files

Share the following .csv results files with the study coordinator (Raechel):

- pathways
- EventCohortCounts
- DistinctEventCohorts
- All Characterization Results files

A secure file transfer protocol can be implemented to facilitate the sharing of the results files as requested by data partners- please reach out to raechel.davis@yale.edu

raechel.davis@yale.edu

Job Opening

COLUMBIA | COLUMBIA UNIVERSITY DEPARTMENT OF BIOMEDICAL INFORMATICS

DBMI Home | News & Events | Research | People | Prospective Students | Academics | Resources

Tenure Track Faculty

#105752

Description

The Department of Biomedical Informatics (DBMI) of Columbia University seeks exceptional junior-level faculty members in the tenure track.

The positions are open to researchers interested in developing and applying informatics theory and achieving tangible benefits to health care and biology. Three particular foci are (1) machine learning for healthcare and health-related data science, (2) health information technology-based interventions to improve health care and the health of individuals and populations, and (3) translational bioinformatics.

Job Opening

Job Details

Database Programmer

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, Windmill Road, Oxford, OX3 7LD

We are seeking to appoint a highly qualified and dedicated Database Programmer to join the Health Data Sciences research group led by Professor Daniel Prieto-Alhambra at the Botnar Research Centre, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), Oxford.

You will join an outstanding, multi-disciplinary and friendly Group of motivated and cutting-edge researchers and to contribute to clinical research by providing technical knowledge, software engineering expertise and data insight.

As a Database Programmer you will Develop new database applications for big clinical data to meet project requirements and deadlines, provide software feedback and carry out software improvement, extension, integration and further development on existing code. You will contribute to the harmonisation, curation, and processing of large clinical datasets and develop code to validate, test, document and maintain database applications. You will also represent the project, team, and the University in collaboration meetings, conferences and at external meetings.

You will have a Degree in computer science, software engineering, health informatics or an equivalent combination of training and professional experience. Proven understanding and experience in one or more RDBMSs and SQL dialects (e.g. PostgreSQL), excellent skills in at least one high level programming language (e.g. Python, C#, C++) and excellent analytical and problem-solving skills with great attention to detail are essential. Experience in common data models (CDMs) and in the extract, transform, and load (ETL) process, knowledge of R and/or RStudio and working experience in a research environment are desirable.

This is a full-time fixed-term appointment for 2 years.

The closing date for this position is 12 noon on Monday 27 February 2023. You will be required to upload a CV and supporting statement as part of your online application.

Contact Person : HR Team, NDRMS
Contact Phone :
Vacancy ID : 163066
Closing Date & Time : 27-Feb-2023 12:00

Summer Internships

General Administration

Epidemiology Graduate Intern

General Administration

OHDA Graduate Intern

General Administration

OHDA Undergraduate Intern

Summer Internships

General Administration

Data Science RWE for R&D Summer Intern

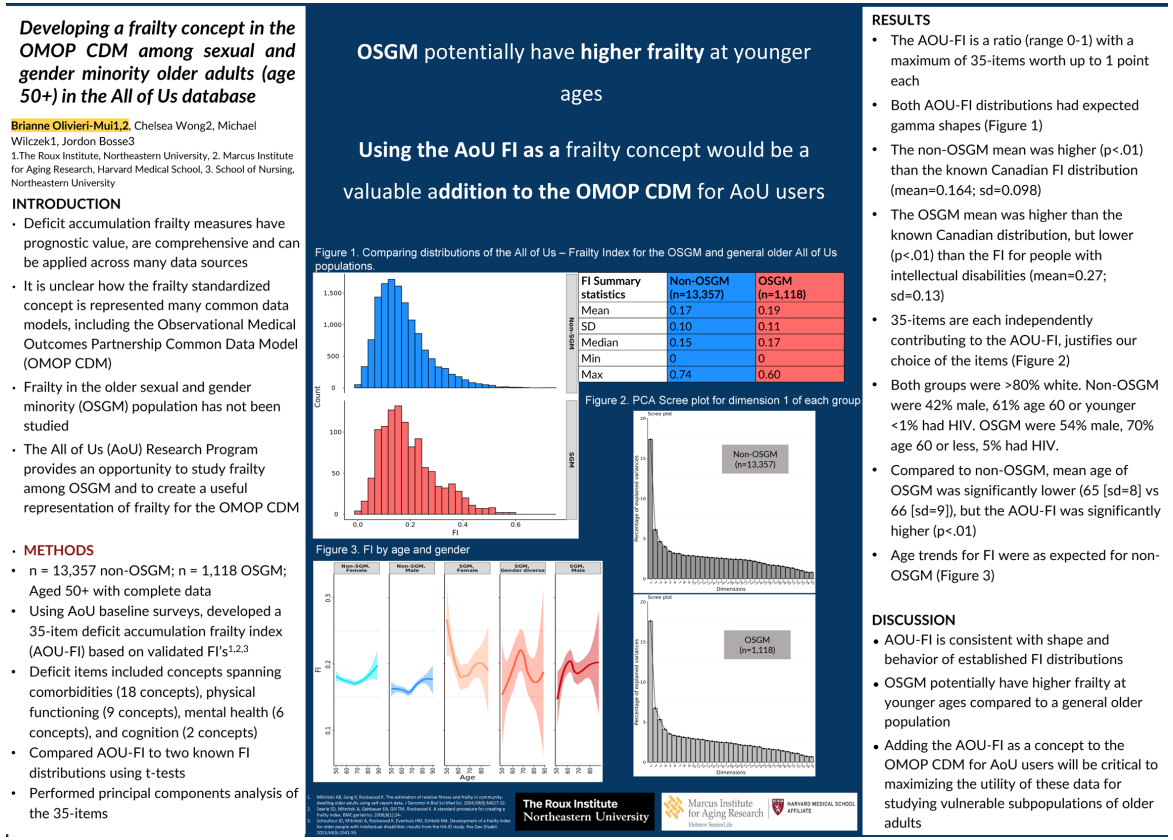
General Administration

Data Science RWE DevCon Summer Intern

#OHDSISocialShowcase This Week

MONDAY ICD10–SNOMED mapping pitfalls: Post-coordinated expressions and concept sets (Sigfried Gold, Tanner Zhang, Richard L. Zhu, Stephanie Hong, Harold P. Lehmann, Davera Gabriel, Tricia Francis, Lisa Eskenazi, Christopher G. Chute)

#OHDSISocialShowcase This Week



Developing a frailty concept in the OMOP CDM among sexual and gender minority older adults (age 50+) in the All of Us database

Brianne Olivieri-Mui1,2, Chelsea Wong2, Michael Wilczek1, Jordon Bosse3
1.The Roux Institute, Northeastern University, 2. Marcus Institute for Aging Research, Harvard Medical School, 3. School of Nursing, Northeastern University

INTRODUCTION

- Deficit accumulation frailty measures have prognostic value, are comprehensive and can be applied across many data sources
- It is unclear how the frailty standardized concept is represented many common data models, including the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)
- Frailty in the older sexual and gender minority (OSGM) population has not been studied
- The All of Us (AoU) Research Program provides an opportunity to study frailty among OSGM and to create a useful representation of frailty for the OMOP CDM

METHODS

- n = 13,357 non-OSGM; n = 1,118 OSGM; Aged 50+ with complete data
- Using AoU baseline surveys, developed a 35-item deficit accumulation frailty index (AOU-FI) based on validated FI's[1,2,3]
- Deficit items included concepts spanning comorbidities (18 concepts), physical functioning (9 concepts), mental health (6 concepts), and cognition (2 concepts)
- Compared AOU-FI to two known FI distributions using t-tests
- Performed principal components analysis of the 35-items

OSGM potentially have **higher frailty** at younger ages

Using the AoU FI as a frailty concept would be a valuable a**ddition to the OMOP CDM** for AoU users

Figure 1. Comparing distributions of the All of Us – Frailty Index for the OSGM and general older All of Us populations.

FI Summary statistics	Non-OSGM (n=13,357)	OSGM (n=1,118)
Mean	0.17	0.19
SD	0.10	0.11
Median	0.15	0.17
Min	0	0
Max	0.74	0.60

Figure 2. PCA Scree plot for dimension 1 of each group

Figure 3. FI by age and gender

RESULTS

- The AOU-FI is a ratio (range 0-1) with a maximum of 35-items worth up to 1 point each
- Both AOU-FI distributions had expected gamma shapes (Figure 1)
- The non-OSGM mean was higher ($p<.01$) than the known Canadian FI distribution (mean=0.164; sd=0.098)
- The OSGM mean was higher than the known Canadian distribution, but lower ($p<.01$) than the FI for people with intellectual disabilities (mean=0.27; sd=0.13)
- 35-items are each independently contributing to the AOU-FI, justifies our choice of the items (Figure 2)
- Both groups were >80% white. Non-OSGM were 42% male, 61% age 60 or younger <1% had HIV. OSGM were 54% male, 70% age 60 or less, 5% had HIV.
- Compared to non-OSGM, mean age of OSGM was significantly lower (65 [sd=8] vs 66 [sd=9]), but the AOU-FI was significantly higher ($p<.01$)
- Age trends for FI were as expected for non-OSGM (Figure 3)

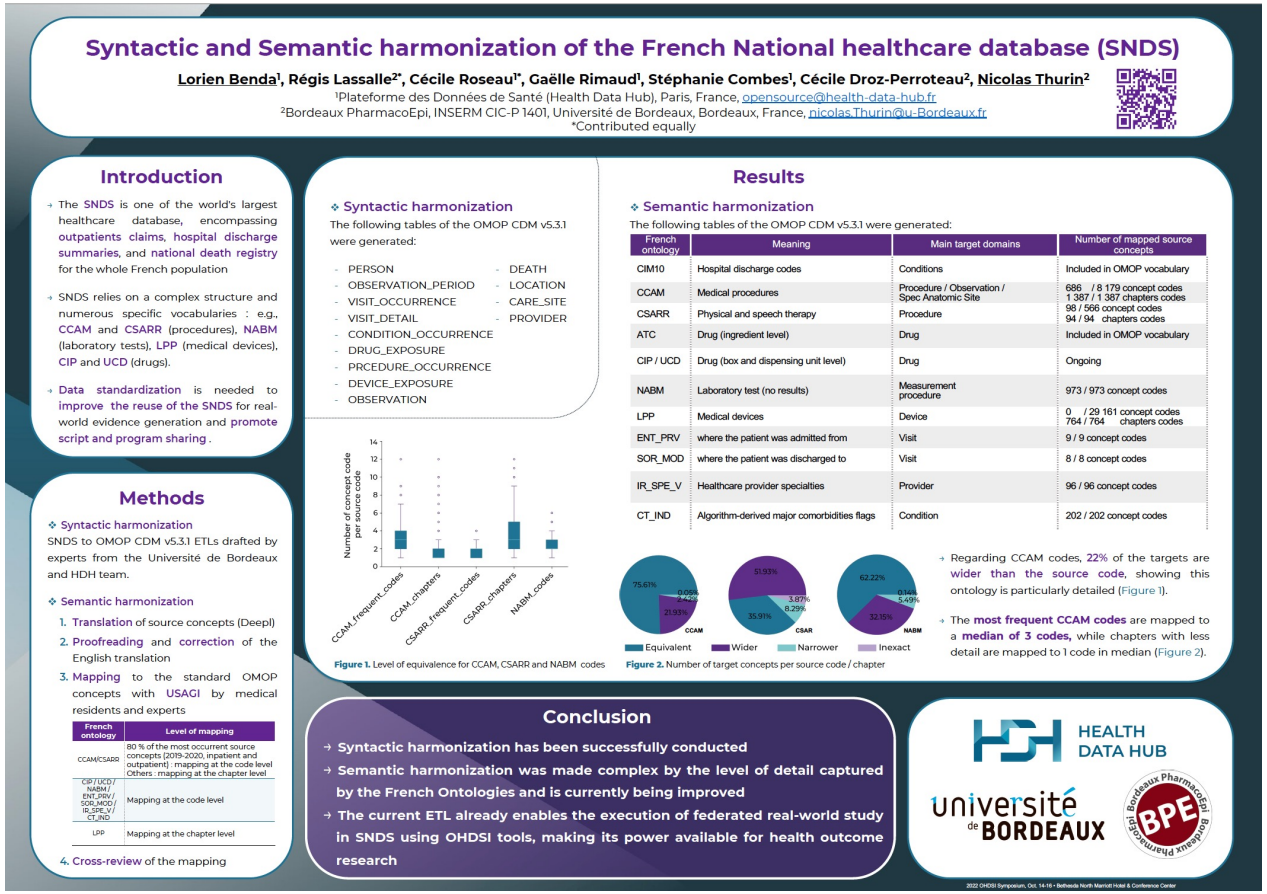
DISCUSSION

- AOU-FI is consistent with shape and behavior of established FI distributions
- OSGM potentially have higher frailty at younger ages compared to a general older population
- Adding the AOU-FI as a concept to the OMOP CDM for AoU users will be critical to maximizing the utility of these data for studying vulnerable subpopulations of older adults

1. Mitnitski AB, Song X, Rockwood K. The estimation of relative fitness and frailty in community-dwelling older adults using self-report data. J Gerontol A Biol Sci Med Sci. 2004;59(6):M627-32.
2. Searle SD, Mitnitski A, Gahbauer EA, Gill TM, Rockwood K. A standard procedure for creating a frailty index. BMC geriatrics. 2008;8(1):24.
3. Schoufour JD, Mitnitski A, Rockwood K, Evenhuis HM, Echteld MA. Development of a frailty index for older people with intellectual disabilities: results from the HA-ID study. Res Dev Disabil. 2013;34(4):1541-51.

TUESDAY **Developing a frailty concept in the OMOP CDM among sexual minority older adults (age 50+) in the All of Us database** (**Brianne Olivieri-Mui**, Chelsea Wong, Michael Wilczek, Jordon Bosse)

#OHDSISocialShowcase This Week



Syntactic and Semantic harmonization of the French National healthcare database (SNDS)

Lorien Benda[1], Régis Lassalle[2*], Cécile Roseau[1*], Gaëlle Rimaud[1], Stéphanie Combes[1], Cécile Droz-Perroteau[2], Nicolas Thurin[2]

[1]Plateforme des Données de Santé (Health Data Hub), Paris, France, opensource@health-data-hub.fr

[2]Bordeaux PharmacoEpi, INSERM CIC-P 1401, Université de Bordeaux, Bordeaux, France, nicolas.Thurin@u-Bordeaux.fr

*Contributed equally

Introduction

- The SNDS is one of the world's largest healthcare database, encompassing outpatients claims, hospital discharge summaries, and national death registry for the whole French population
- SNDS relies on a complex structure and numerous specific vocabularies : e.g., CCAM and CSARR (procedures), NABM (laboratory tests), LPP (medical devices), CIP and UCD (drugs).
- Data standardization is needed to improve the reuse of the SNDS for real-world evidence generation and promote script and program sharing .

Methods

❖ Syntactic harmonization

SNDS to OMOP CDM v5.3.1 ETLs drafted by experts from the Université de Bordeaux and HDH team.

❖ Semantic harmonization

1. Translation of source concepts (Deepl)
2. Proofreading and correction of the English translation
3. Mapping to the standard OMOP concepts with USAGI by medical residents and experts

French ontology	Level of mapping
CCAM/CSARR	80 % of the most occurrent source concepts (2019-2020, inpatient and outpatient) : mapping at the code level Others : mapping at the chapter level
CIP / UCD / NABM / ENT_PRV / SOR_MOD / IR_SPE_V / CT_IND	Mapping at the code level
LPP	Mapping at the chapter level

4. Cross-review of the mapping

Syntactic harmonization

The following tables of the OMOP CDM v5.3.1 were generated:

- PERSON
- OBSERVATION_PERIOD
- VISIT_OCCURRENCE
- VISIT_DETAIL
- CONDITION_OCCURRENCE
- DRUG_EXPOSURE
- PRCEDURE_OCCURRENCE
- DEVICE_EXPOSURE
- OBSERVATION
- DEATH
- LOCATION
- CARE_SITE
- PROVIDER

Figure 1. Level of equivalence for CCAM, CSARR and NABM codes

Results

❖ Semantic harmonization

The following tables of the OMOP CDM v5.3.1 were generated:

French ontology	Meaning	Main target domains	Number of mapped source concepts
CIM10	Hospital discharge codes	Conditions	Included in OMOP vocabulary
CCAM	Medical procedures	Procedure / Observation / Spec Anatomic Site	686 / 8 179 concept codes 1 387 / 1 387 chapters codes
CSARR	Physical and speech therapy	Procedure	98 / 566 concept codes 94 / 94 chapters codes
ATC	Drug (ingredient level)	Drug	Included in OMOP vocabulary
CIP / UCD	Drug (box and dispensing unit level)	Drug	Ongoing
NABM	Laboratory test (no results)	Measurement procedure	973 / 973 concept codes
LPP	Medical devices	Device	0 / 29 161 concept codes 764 / 764 chapters codes
ENT_PRV	where the patient was admitted from	Visit	9 / 9 concept codes
SOR_MOD	where the patient was discharged to	Visit	8 / 8 concept codes
IR_SPE_V	Healthcare provider specialties	Provider	96 / 96 concept codes
CT_IND	Algorithm-derived major comorbidities flags	Condition	202 / 202 concept codes

Figure 2. Number of target concepts per source code / chapter

- Regarding CCAM codes, 22% of the targets are wider than the source code, showing this ontology is particularly detailed (Figure 1).
- The most frequent CCAM codes are mapped to a median of 3 codes, while chapters with less detail are mapped to 1 code in median (Figure 2).

Conclusion

- Syntactic harmonization has been successfully conducted
- Semantic harmonization was made complex by the level of detail captured by the French Ontologies and is currently being improved
- The current ETL already enables the execution of federated real-world study in SNDS using OHDSI tools, making its power available for health outcome research

HEALTH DATA HUB

université de BORDEAUX

Bordeaux PharmacoEpi

2022 OHDSI Symposium, Oct. 14-16 • Bethesda North Marriott Hotel & Conference Center

WEDNESDAY

Syntactic and Semantic Harmonization of the French National Healthcare Database (SNDS) (Lorien Benda, Regis Lassalle, Cecile Roseau, Stephanie Combes, Cecile Droz-Perroteau, Nicolas Thurin)

#OHDSISocialShowcase This Week

Using Data Augmentation for NER-RE Joint-Learning Tasks for Clinical History Information Extraction

Xiaodong Zhu[1], Miao Chen[2], Daniel Slaughter[1], Elizabeth L. Lyon[1], Pallavi Misra[1] and Michael Biorn[1]
[1]Labcorp, Princeton, NJ; [2]Microsoft Corporation, Redmond, WA

Introduction

In the clinical trial and drug development industry, patient clinical history information is critical in helping determine the eligibility of a patient to be enrolled in a clinical trial. Unfortunately, a great deal of patient history is in text format, which means we need to employ natural language processing (NLP) methods to extract needed structured information. We have previously constructed a transformer BERT-based, joint-learning model for NER and RE tasks, which successfully extracted information from clinical trial protocols.[1] However, we face a new challenge that we do not possess sufficiently labeled data. In fact, lacking training data is a major obstacle for applying deep learning techniques in the medical domain.[2]

Recently, multiple methods for data augmentation have been investigated, including both rule-based and neural network-based methods.[3-5] Dai and Adel's study applied four different ways of transforming text for text augmentation purposes: Label-wise Token Replacement (LwTR), Shuffle within Segments (SiS), Synonym Replacement (SR) and Mention Replacement (MR)[6] and yielded boosted performance. Kang et al. used synonyms from UMLS for data augmentation.[7] Inspired by these studies, we applied three transformations: LwTR, SiS and MR to generate the augmented data. We performed the entity replacement not only with the UMLS synonyms, but also used the broader concepts and narrower concepts. Our results demonstrated that such data augmentation can dramatically improve the generalization of the NER-RE joint-learning model.

Table 1A. Data Counts for NER Task

Entity Type	Train	Test
Cancer	348	82
Condition	262	35
Other_disease	111	17
Differential_consideration	104	22
Anatomic_location	43	13
Biomolecule	35	6
Qualifier_modifier	30	6
Procedure	26	9
Temporal_constraint	22	2
Remission_condition	17	6
Negation_cue	13	5
Stage	8	2
ICD	5	2

Table 1B. Data Counts for RE Task

Relation Type	Train	Test
Has_consideration	143	31
Modified_by	95	20
Is_located	45	15
Has_history	31	6
Has_temMea	28	3
Is_negated	13	5

Results

As shown in Table 2, training with the augmented data improved both the NER and RE tasks. NER performance was improved from F1 score of 0.71 to 0.75. Interestingly, although no transformation was directly applied to the relation, we found RE performance was improved from F1 score of 0.34 to 0.44. This was not surprising though, as in the joint-learning model, RE and NER tasks shared the hidden representation and the loss was optimized towards both tasks. Thus, data augmentation designed for NER helps the RE task.

Table 2. NER and RE Task Performance from Models Trained with Different Data Sets

Data	Tasks	Precision	Recall	F1
Original	NER performance	0.69	0.77	0.71
Original + Augmentation	NER performance	0.74	0.77	0.75
Original	RE performance	0.32	0.42	0.34
Original + Augmentation	RE performance	0.51	0.375	0.44

Methods

In order to obtain high-quality labeled data, we sought help from subject matter experts (SME) to annotate a small set of clinical history text. The SMEs looked for and highlighted oncology-related entities. For the current work, we used 13 named entity types and 6 types of relations between entities. In most cases, one record contained multiple entities and relations. Figure 1 shows the entity and relation counts.

We obtained 432 clinical history records in total, which served as seed data for the data augmentation tasks. We split the data into two parts, with one part of 362 records as a training set and a seed set for augmentation, with the remaining 70 records as a test set. Table 1 shows the entity and relation types and counts (before augmentation).

We applied LwTR (Figure 1B), SiS (Figure 1C) and MR (Figure 1D) transformations for data augmentation by following the methods described by Dai and Adel.[8] Briefly, for each token/segment/entity, we first sampled from the binomial distribution with p = 0.5 to determine whether the transformation should be performed. For LwTR, tokens with the same entity labels were sampled randomly from the training set and used as replacements. For SiS, new sentences were generated by randomly shuffling tokens within a sentence segment. For MR, entities with the same type were sampled randomly from the training set and used as the replacement. Each of these transformations generated 362 new records.

To replace a named entity with a related concept from UMLS, we focused on four entity types: Condition, Cancer, Other Disease and Anatomic Location. For each entity within these four categories, we sampled from a binomial distribution with a fixed p value to determine whether it should be replaced. We obtained three sets of data with p = 0.5, 0.6 and 0.7, respectively. For each set, the UMLS ontology was used to retrieve the narrower concepts (Figure 1E), broader concepts (Figure 1F) and the synonyms (Figure 1G). If no related concepts could be found from UMLS, the replacement was skipped. In total we obtained 1,319 records using UMLS concepts.

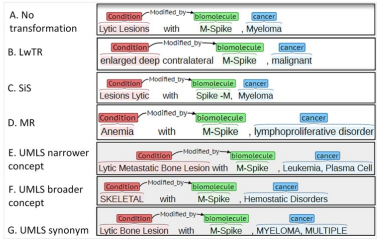


Figure 1. Data augmentation with different methods. A. Original records, no transformation. **B.** LwTR (Label-wise Token Replacement). **C.** SiS (Shuffle within Segments). **D.** MR (Mention Replacement). Named entities were replaced. **E-F.** Entities were replaced by the randomly selected narrower concept in UMLS. Note that 'M-Spike' was not replaced as only entities of Condition, Cancer, Other Disease and Anatomic Location were processed.

Conclusions

In real-world situations, it is very difficult and costly to obtain a large annotated clinical text data set. This study demonstrated that data augmentation can improve both the NER and RE tasks for information extracted from patient clinical history. The data augmentation methods used here were rule-based transformation of the original training data, and rule-based replacement of terms from the UMLS ontology. For future work, it would be interesting to experiment with neural network-based data augmentation methods, as well as to evaluate how the extraction results impact downstream business tasks such as cohort selection.

References

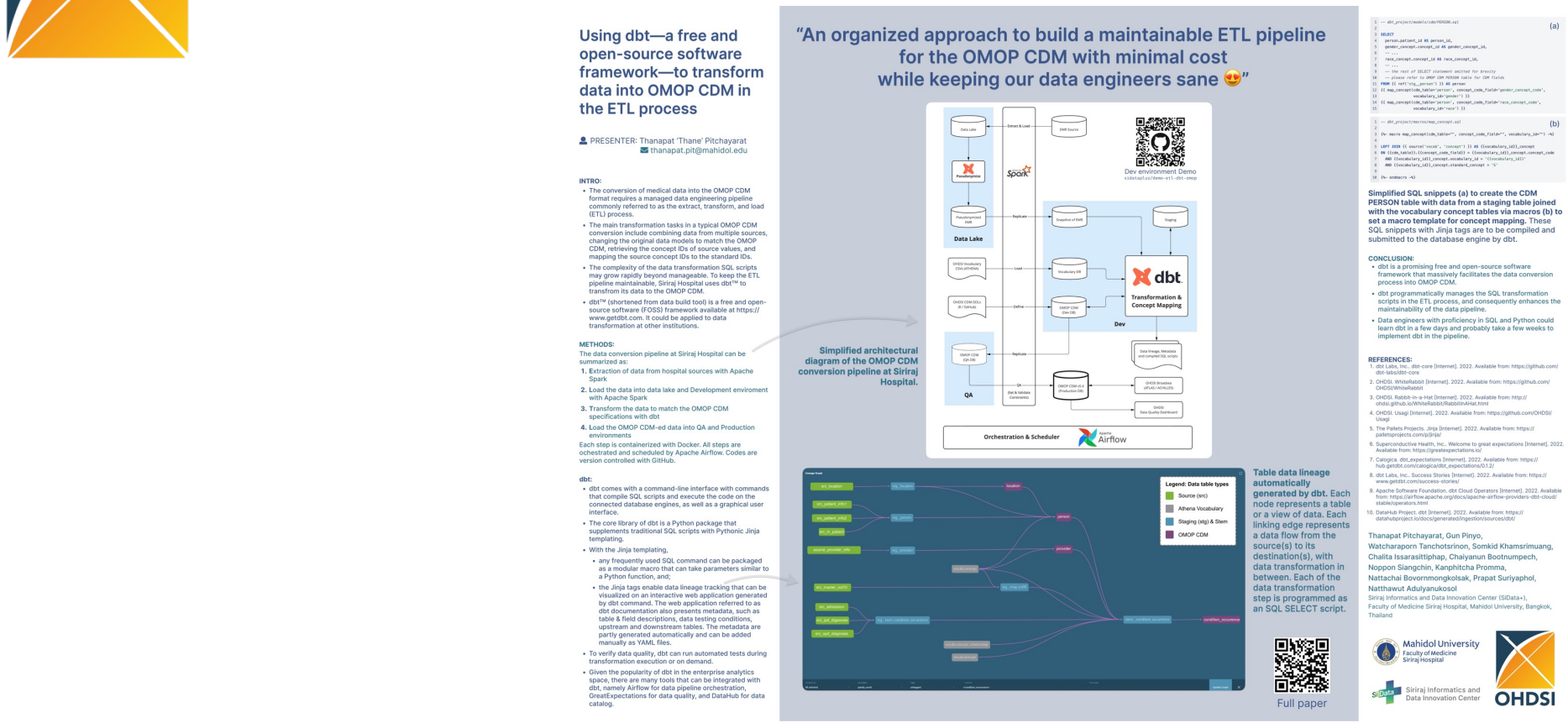
1. Chen M, Lan G, Du F, Lobanov VS, Joint Learning with Pre-trained Transformer on Named Entity Recognition and Relation Extraction Tasks for Clinical Analytics. CLINICALNLP; 2020.1.
2. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. npj Digital Medicine. 2019;2(1):4
3. Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, et al. A Survey of Data Augmentation Approaches for NLP, ArXiv. 2021;abs/2105.03075.
4. Chen J, Tam D, Raffel C, Bansal M, Yang D. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. ArXiv. 2021;abs/2106.07499.
5. Shorten C, Khoshgoftaar TM, Furht B. Text Data Augmentation for Deep Learning. Journal of Big Data. 2021;8(1):101.
6. Dai X, Adel H. An Analysis of Simple Data Augmentation for Named Entity Recognition. ArXiv. 2020;abs/2010.11683.
7. Kang T, Perotte A, Tang Y, Ta C, Weng C. UMLS-based data augmentation for natural language processing of clinical research literature. J Am Med Inform Assoc. 2021;28(4):812-23.

Presented at OHDSI Symposium 2022

labcorp

THURSDAY Using data augmentation for NER-RE joint learning tasks for clinical history information extraction (Xiaodong Zhu, Miao Chen, Daniel Slaughter, Elizabeth Lyon, Pallavi Misra, Michael Biorn)

#OHDSISocialShowcase This Week



FRIDAY

Using dbt – a free and open-source software – to transform data into OMOP CDM in the ETL process (Thanapat Pitchayarat, Gun Pinyo, Watcharaporn Tanchotsrinon, Somkid Khamsrimuang, Chalita Issarasittiphap, Chaiyanun Bootnumpech, Noppon Siangchin, Kanphitcha Promma, Nattachai Bovornmongkolsak, Prapat Suriyaphol, Natthawut Adulyanukosol)

Where Are We Going?

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

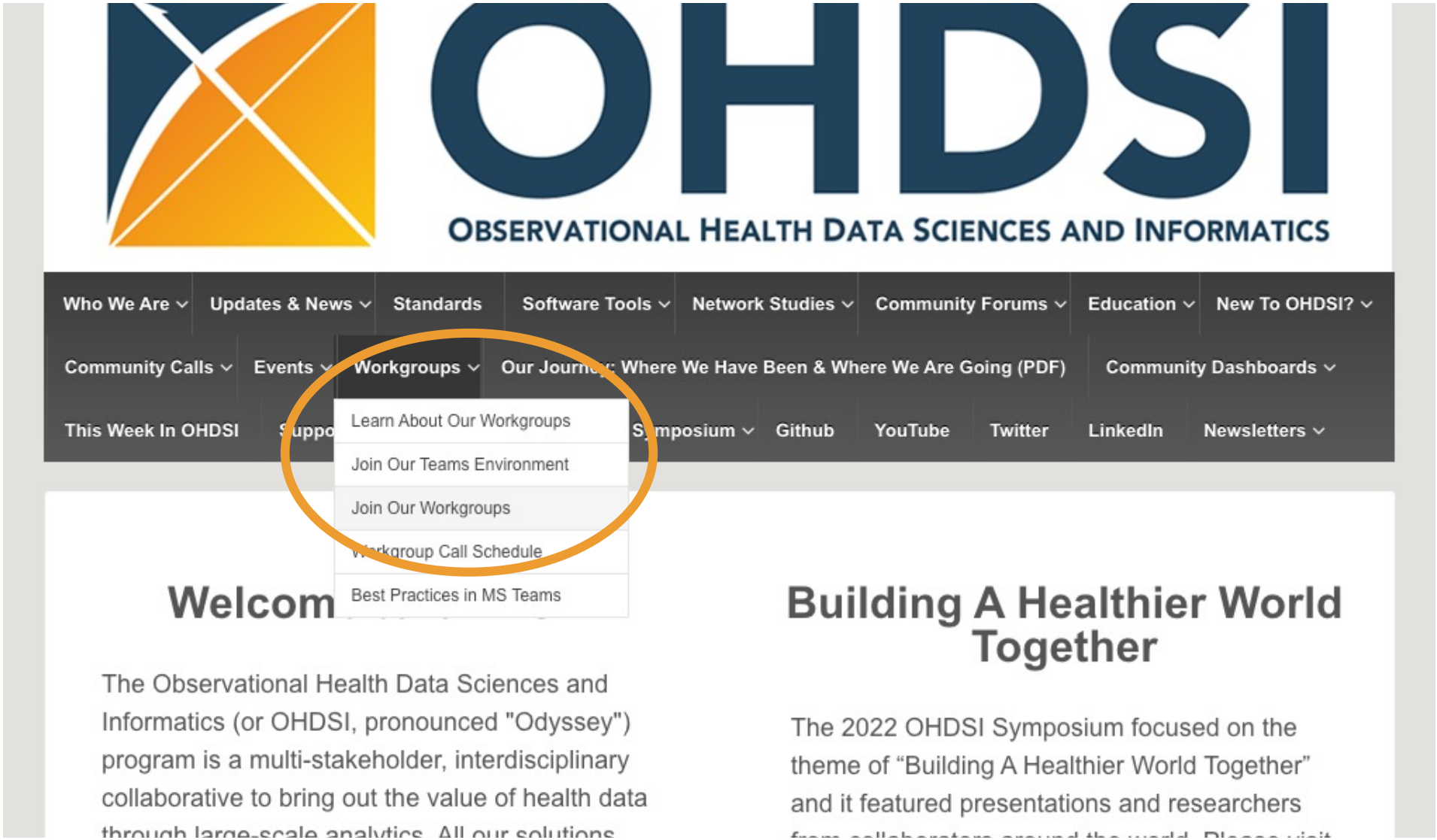
Three Stages of The Journey

Where Have We Been?

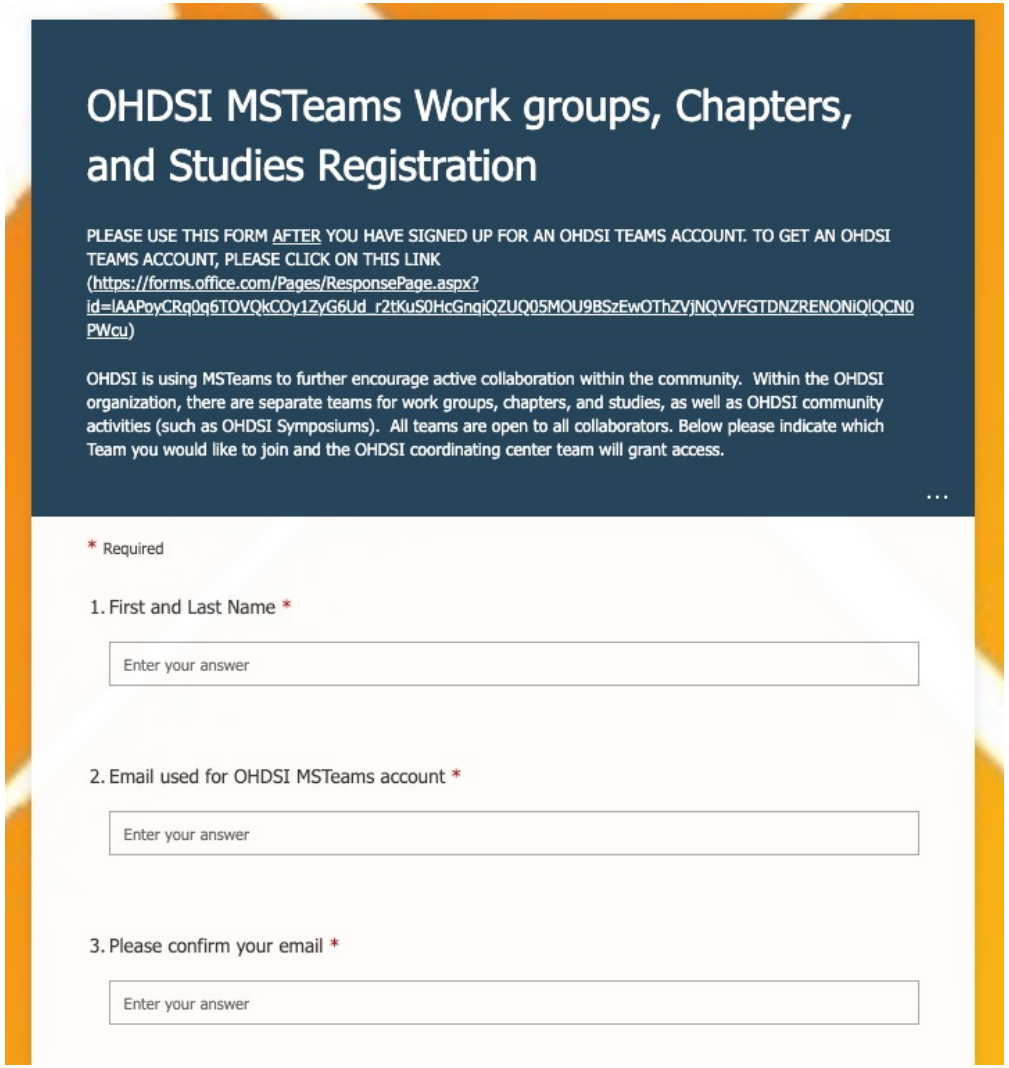
Where Are We Now?

Where Are We Going?

How To Join The Workgroups



How To Join The Workgroups



OHDSI MSTeams Work groups, Chapters, and Studies Registration

PLEASE USE THIS FORM AFTER YOU HAVE SIGNED UP FOR AN OHDSI TEAMS ACCOUNT. TO GET AN OHDSI TEAMS ACCOUNT, PLEASE CLICK ON THIS LINK (https://forms.office.com/Pages/ResponsePage.aspx?id=lAAPoyCRq0q6TOVQkCOy1ZyG6Ud_r2tKuS0HcGnqiQZUQ05MOU9BSzEwOThZVjNQVVFGTDNZRENONiQlQCN0PWcu)

OHDSI is using MSTeams to further encourage active collaboration within the community. Within the OHDSI organization, there are separate teams for work groups, chapters, and studies, as well as OHDSI community activities (such as OHDSI Symposiums). All teams are open to all collaborators. Below please indicate which Team you would like to join and the OHDSI coordinating center team will grant access.

...

* Required

1. First and Last Name *

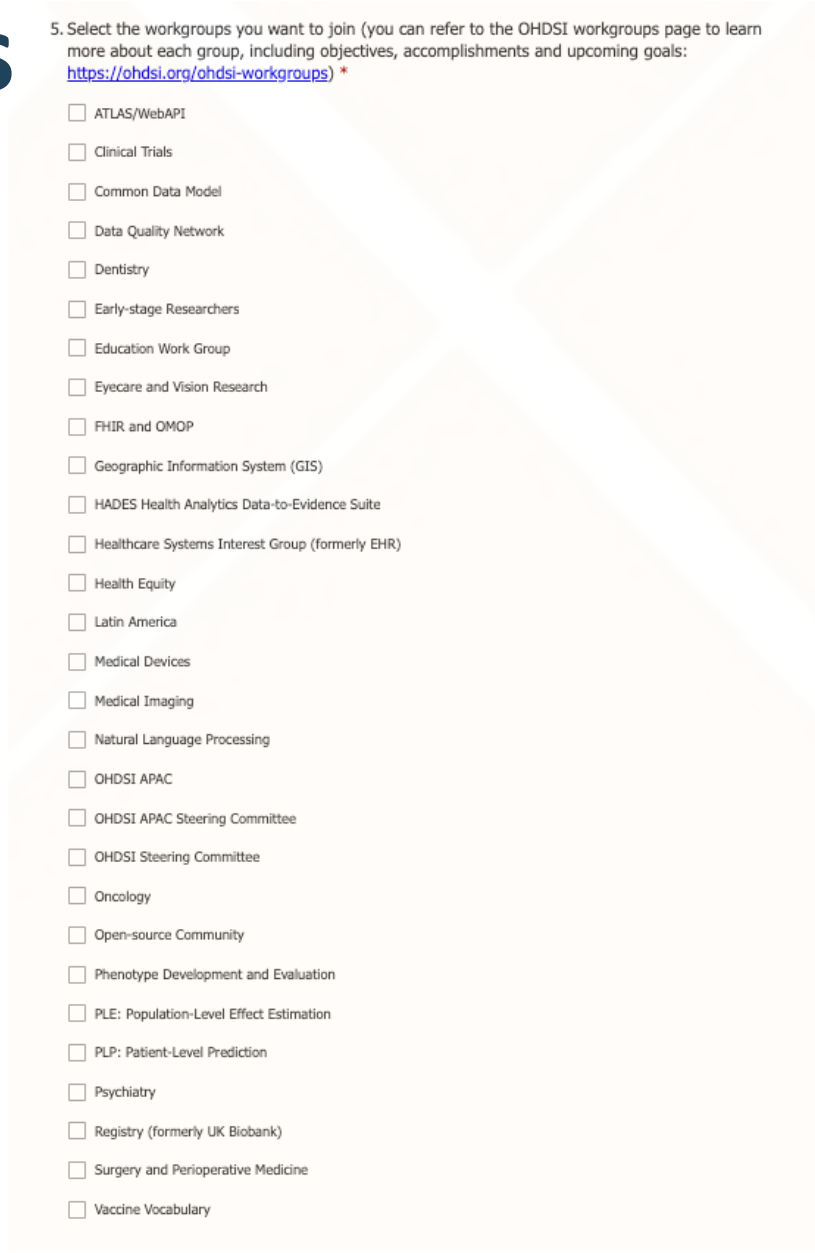
Enter your answer

2. Email used for OHDSI MSTeams account *

Enter your answer

3. Please confirm your email *

Enter your answer



5. Select the workgroups you want to join (you can refer to the OHDSI workgroups page to learn more about each group, including objectives, accomplishments and upcoming goals: https://ohdsi.org/ohdsi-workgroups) *

- [ ] ATLAS/WebAPI
- [ ] Clinical Trials
- [ ] Common Data Model
- [ ] Data Quality Network
- [ ] Dentistry
- [ ] Early-stage Researchers
- [ ] Education Work Group
- [ ] Eyecare and Vision Research
- [ ] FHIR and OMOP
- [ ] Geographic Information System (GIS)
- [ ] HADES Health Analytics Data-to-Evidence Suite
- [ ] Healthcare Systems Interest Group (formerly EHR)
- [ ] Health Equity
- [ ] Latin America
- [ ] Medical Devices
- [ ] Medical Imaging
- [ ] Natural Language Processing
- [ ] OHDSI APAC
- [ ] OHDSI APAC Steering Committee
- [ ] OHDSI Steering Committee
- [ ] Oncology
- [ ] Open-source Community
- [ ] Phenotype Development and Evaluation
- [ ] PLE: Population-Level Effect Estimation
- [ ] PLP: Patient-Level Prediction
- [ ] Psychiatry
- [ ] Registry (formerly UK Biobank)
- [ ] Surgery and Perioperative Medicine
- [ ] Vaccine Vocabulary

OHDSI China

Lei Liu

China Chapter Purpose

OHDSI China WG exists to promote OHDSI methodologies and collaboration among healthcare providers, biomedical research institutions, and industry in China and connect Chinese healthcare research to the global community, providing training and workshops and some technical support to facilitate local adaptation of OHDSI strategy.

China Chapter 2023 Objectives and Key Results

Objective 1: Promote OHDSI strategy and methodology in China

Key results:

- Organize at least three online training workshops in China Timeline: 1-2Q2023
- Continue monthly online invited presentation Timeline: 1-4Q2023
- Organize OHDS China Annual Symposium Timeline: 4Q2023

Objective 2: Create collaboration activities that encourage collaborative research among healthcare institutions in China

Key results:

- Implement data mapping and ETL process for at least three hospitals using OMOP model data Timeline: 1-3Q2023
- Initiate 1-2 collaborative studies on certain diseases from at least two healthcare institutions using OHDSI strategy Timeline: 3-4Q2023

Natural Language Processing

Hua Xu

NLP WG Purpose

The NLP WG exists to promote the use of textual information in electronic health records (EHRs) by developing software tools and methods to represent and utilize textual data thereby facilitating the generation of evidence for observational studies.

NLP 2023 Objectives and Key Results

Objective 1: Knowledge dissemination - contribute a chapter on NLP in the Book of OHDSI

Key results

1. Deliver the initial draft of the chapter; Timeline: 1-3Q2023

Objective 2: ETL for textual data normalization

Key results

1. Continue developing the tool to normalize “note_type”, release version 1.0; Timeline: 1-4Q2023
2. Discuss and revise the schema for relation representation; release an initial draft, Timeline: 1-2Q2023

Objective 3: Conduct at least two multi-site clinical studies that utilize both structured and textual data

Key results

1. Descriptive study on Social determinants of health; Timeline: 1-4Q2023
2. Identify delirium episodes from clinical notes; Timeline: 1-4Q2023

ATLAS Working Group

OHDSI Community Call

14 Feb 2023

Mission

- The ATLAS workgroup will provide a forum for the OHDSI community of developers that are interested in improving the open-source software solutions: ATLAS & WebAPI. These tools aim to provide capabilities to design standardized analytics to execute on the OMOP Common Data Model.

2022 OKRs: Atlas Working Group

1. Objective: Release Atlas/WebAPI v2.13
   - Complete the development and testing of existing issues
   - Target delivery: 1Q2023

2. Objective: After v2.13 release, establish and maintain a robust software development plan for ATLAS moving forward
   - Identify candidate features for future ATLAS releases through active solicitation from the OHDSI community and ATLAS users
   - Candidate features for the next release (i.e Atlas v2.14) will be prioritized and selected by product owners (Patrick Ryan/Greg Klebanov)
   - Project Managers (Anthony Sena/Alexey Manoylenko) will identify developers to lead the design and implementation.
   - Developers will produce design specification document for each feature and the Technical Lead (Chris Knoll) will review and approve prior to development.
   - Once all features are implemented, Project Managers will coordinate testing across 3 independent sites: Odysseus, Janssen, and public OHDSI-test instance
   - Once Product owners approve a release, Project Managers will then communicate release and documentation to the OHDSI community

2023 Meeting Schedule

- Monthly ATLAS WG Meetings 1st Monday of the month at 9AM EST with the following aims
  - Highlight new features being built in the community
  - Provide updates on upcoming releases

- Weekly Developer Meetings- Tuesday, 8:30AM EST
  - Review open pull requests for Atlas/WebAPI
  - Triage/Review issues for upcoming releases

Registry Workgroup

Tina Parciak

Purpose of the Registry WG

- Registries contain valuable real-world data whose potential could be increased by transforming it to the OMOP common data model.
- Registry data might face additional challenges compared to data coming from electronic health care records.
  - For example, registry data often consists of data that does not come from EHR (alone) and uses other standards, if any, than those that are widely used in EHR.
- As a group we aim to identify these challenges and align approaches to transform registry data to the OMOP CDM.

Purpose of the Registry WG

- Reasons of transforming registry data to OMOP CDM:
  - Become part of projects / networks (e.g. EHDEN)
  - (Inter)national research / collaboration
  - Enrichment or validation of own studies with (linked) OMOPed data
  - Automation and lowering physician burden for data collection
  - Make registry data more FAIR
  - Create standardized analysis across databases in a more transparent way with more reliable results
  - Privacy by design

Members' voices

Purpose of the Registry WG

- Registries contain valuable real-world data whose potential could be increased by transforming it to the OMOP common data model.
- Registry data might face additional challenges compared to data coming from electronic health care records.
  - For example, registry data often consists of data that does not come from EHR (alone) and uses other standards, if any, than those that are widely used in EHR.
- As a group we aim to identify these challenges and align approaches to transform registry data to the OMOP CDM.

2022 Achievements of the Registry WG

- We successfully constituted the Registry WG out of the former UK Biobank WG and gathered interested, motivated OMOP collaborators.
- Kick-started with a survey to get the first impressions of the group's members, focus and ideas.
- We had several talks from different registries who transformed their data to the OMOP CDM and discussed the challenges.

Get to know the Registry WG - Survey

“What is a registry?“

"a collection of standardized information about a group of patients who share a disease or experience". (NIH)

- “A **structured data set** with specific variables collected for a specific set of patients.”
- “organised collection of **defined and standardised health data** on a **specific disease or patient population**”
- “A **centralization of clinical data** about patients across multipole locations with a **standardized data collection protocol** and generally a list of questions to answer.”
- “**long-term data collection system** on a group of person with a specific disease or condition and established for a specific purpose”
- “**Systematic collection of health information** about individuals and *regulated by law*.”
- “high quality platform”
- “Extracted and summarized **EHR data**”

Get to know the Registry WG - Survey

14. Is data collection currently ongoing?

More Details

Yes, but it has a defined end point	2
Yes, and it will continue without ...	5
No	2
Unsure	0

2. How would you describe your role associated with this registry and/or the OMOP CDM best?

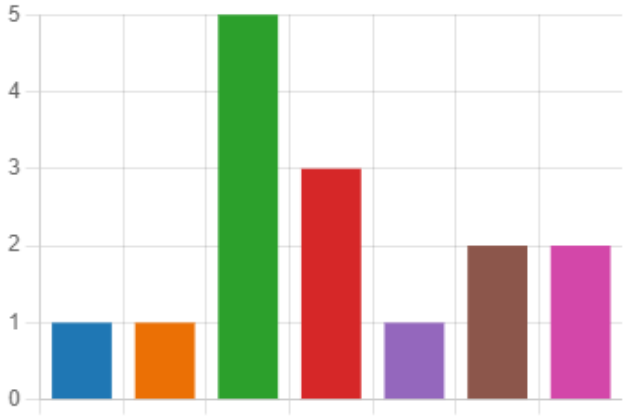
More Details

Data custodian/owner (of regist...	1
Technical expert (e.g. responsibl...	6
Researcher (e.g. uses registry da...	4
Third-party (e.g. industry partne...	1
Other	4

15. What describes the type of your registry best? (multiple answers possible)

More Details

Clinical trial	1
Post-authorization safety study	1
Clinician-driven registry (dedicat...	5
Patient-driven registry	3
Clinician or patient-driven regist...	1
Registry consisting of linked dat...	2
Other	2



2022 Achievements of the Registry WG

- We successfully constituted the Registry WG out of the former UK Biobank WG and gathered interested, motivated OMOP collaborators.
- Kick-started with a survey to get the first impressions of the group's members, focus and ideas.
- We had several talks from different registries who transformed their data to the OMOP CDM and discussed the challenges.

OKR 2023 (Q1-2) of the Registry WG

- Our main objective for the first half year 2023 is to establish a regular meeting routine of at least a meeting/presentation every 2 months.
- Move ahead with the f ons from the presented ex the OMOP common data r ing.
  - → Mapping conventions
  - → Support for transforn

5. What are your preferences for the frequency of the Registry WG meeting?

More Details

Bi-weekly	0
Once a month	8
Once every 3 month	3
I don't have a preference.	3
Other	0

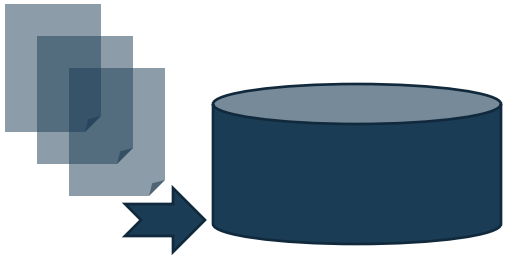
OKR 2023 (Q1-2) of the Registry WG

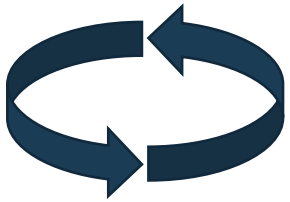
- Our main objective for the first half year 2023 is to establish a regular meeting routine of at least a meeting/presentation every 2 months.
- Move ahead with the following objective of drawing up conclusions from the presented experiences of / challenges in transforming registry data to the OMOP common data model.
  - → Mapping conventions for alignment (2023/2024)
  - → Support for transformations within the workgroup

Common Data Model Vocabulary Subgroup

Christian Reich, Anna Ostropolets, Michael Kallfelz

Workgroup Purpose





- To support the OHDSI community by providing well curated standardized ontologies as well as robust provenance of concept mapping.
- To provide a place for the community to discuss current developments, vocabulary needs or bring forward complaints other than the forums.
- To connect Workgroups with each other to help addressing their specific vocabulary needs as well as to create synergies.

2023 Objectives

& Key Results

Run review sessions for pending changes in specific vocabularies and their build processes

Three vocabularies have been covered

Provide guidance and instructions for community members' contributions

Instructions for providing content input and documentation of the build process available

Implement a standard approach for mapping contribution

Mapping input tables based on SSSOM are defined

Logic available to store, merge and translate mappings

OHDSI Dentistry Workgroup Objectives and Key Results (OKR)

Lead: Robert Koski

Work Group Mission

To understand how dentistry can leverage observational research to improve oral health outcomes and further investigate the links between oral health and systemic disease.

Some of the ways we plan to support this mission:

- Increased adoption of observational research in dentistry
- OMOPification of dental datasets
- Observational research studies driven by use cases

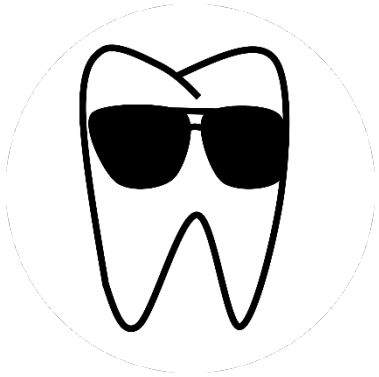
Objectives and Key Results

1. **Expand the workgroup**
    a. Recruit a co-lead
    b. Recruit at least 5 regularly attending members (up from 3!)
2. **Increase our understanding of observational research in dentistry and promote awareness of observational research in the dental community**
    a. Complete scoping review on observational research in dentistry and submit for publication
    b. Form partnership with ADA SCDI to begin work on a dental standard for common data models
3. **Increase the capabilities of the dental community to conduct observational research**
    a. Develop 5 dental use cases
    b. Find dental dataset and map to the OMOP-CDM
    c. Explore the role of dental radiology/dental imaging in observational research

Dentistry Work Group

WG Meetings

Thursdays at 7PM ET on MS Teams



APAC Steering Group

Mui Van Zandt

APAC Steering Group Purpose

APAC Steering Group exists to support the APAC community in collaboratively generating the evidence that promotes better health decisions and better care, by organizing and guiding collaborative activities and facilitating communications across the community.

OHDSI APAC 2023 Objectives and Key Results

Objective 1: Build research expertise and collaboration amongst the different chapters through publication

Key results:

1. Establish a scientific forum where researchers can collaborate; Timeline: 1Q2023
2. 4 publications; Timeline: 4Q2023

Objective 2: Create an APAC training program to expand reach to the general community

Key results:

1. Determine appropriate structure and format of the training program; Timeline: 1Q2023
2. Host at least 6 training sessions throughout the year; Timeline: 4Q2023
3. Develop and implement training curriculum and logistics; Timeline: 3Q2023
4. Launch APAC training program POC; Timeline: 4Q2023

Objective 3: Create collaboration activities that encourage collaborative generation and dissemination of the evidence that promotes better health decisions and better care

1. Host 1 APAC symposium; Timeline: 3Q2023

OHDSI GIS WG

Leads: Andrew Williams, Kyle Zollo-Venecek, Robert Miller

Goal 1 (Q1 2023)

Finalize development of the schema for the optional GIS OMOP module that consists of a universal schema for storing geospatial data. This module supports standardized query execution on a combination of place-related and person-level data using spatiotemporal relations.

- Present proposal to the OHDSI CDM Workgroup and integrate feedback.
- Once finalized, provide documentation on schema and mechanisms for implementation

Goal 2 (Q1 2023)

Expand the corpus of metadata to include new data sources to advance development of automated retrieval, ingestion, and transformation of additional data sources into the module.

- Create metadata for new datasets, such as: Child Opportunity Index data, Area Deprivation Index, IDSR Disease surveillance, etc.
- Expand metadata and related functionality to enable automated retrieval from APIs.

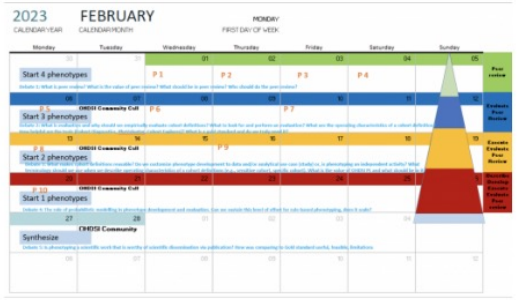
Goal 3 (Q1 2023)

Develop and test **tooling to integrate OMOP clinical data with standardized place-related data**

- Test querying capacity for joining place-related and person-level data by defining cohort definition that includes both clinical and place-related data.
- Engage with our own stats group to create the first draft of a fully specified analytics plan for one of our established topic goals.
- Explore avenues of integration with existing OHDSI tooling, e.g., ATLAS/WebAPI; HADES

New Phenotype Phebruary Homepage

Phenotype Phebruary 2023: How To Join The Effort

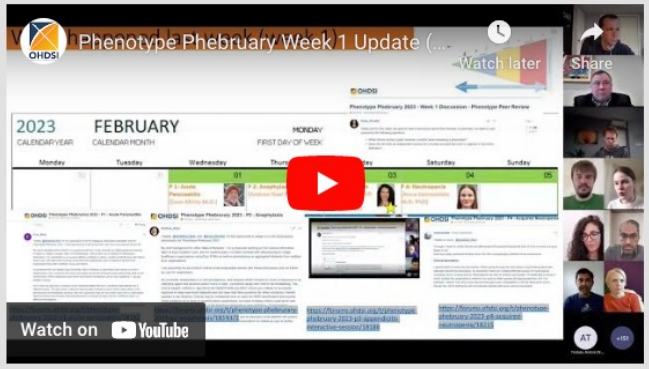


The schedule to the left lists the phenotypes that will be investigated throughout the month, along with the respective leads and reviewers. Check for updates to this graphic as more people join the effort. The graphic to the right highlights the four debates/discussions around phenotyping that are happening this month. Please use the forum links below to join any of these activities.

"Phenotype Phebruary" is a community-wide initiative to both develop and evaluate phenotypes for health outcomes that could be investigated by the community.

This is the second year of Phenotype Phebruary in the OHDSI community (look back at Year 1 here). It was introduced during the Jan. 31 community call (watch here), and will go on throughout the month. This year, the leadership team of **Gowtham Rao** and **Azza Shoaibi** helped identify 10 phenotypes that are being investigated throughout the month. If you would like to join the discussions around any of the phenotypes, please visit the appropriate links below, which will take you to the proper threads on the OHDSI forums.

Week 1 Update



Join Our Community Efforts Around Any Of These Phenotypes

(when phenotype threads get initiated, they will be added to the chart below)

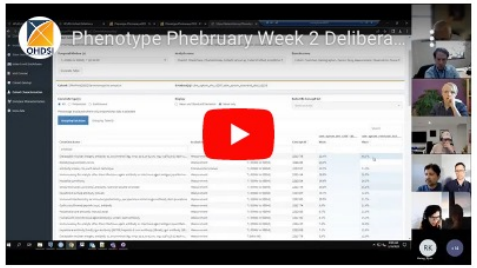
Announcements and Meeting/Workshop Links	Acute Pancreatitis	Anaphylaxis	Appendicitis
Acquired Neutropenia	Systemic Lupus Erythematosus	Acute Hepatic Failure	Idiopathic Inflammatory Myopathies
Parkinson's Disease	ST Elevation Myocardial Infarction	Neonatal Hypoxic Ischemic Encephalopathy	Neurofibromatosis type 1 with Optical Pathway Glioma

Join Our Community Discussions Around These Phenotype Phebruary Topics

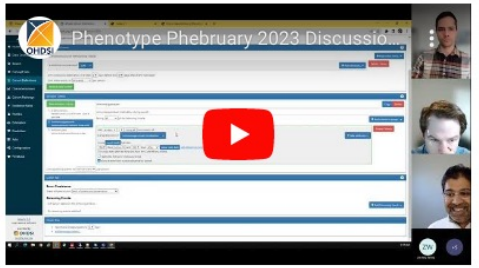
(when phenotype threads get initiated, they will be added to the chart below)

Phenotype Peer Review	Chart review gold standard validation vs innovative methods like PheValuator
What makes cohort definitions reusable, and what is the value of the OHDSI Phenotype Library? What should be in it?	The role of probabilistic modeling in phenotype development and evaluation

Phenotype Phebruary Videos



(Feb. 10) Week 2 of Phenotype Phebruary concluded with this OHDSI Phenotype Development and Evaluation workgroup meeting. In this session, the workgroup assigned leads to each phenotype that are



(Feb. 8) Christopher Mecoli, MD, and team demonstrated progress in the development of a cohort definition for Inflammatory Dermatomyositis at Johns Hopkins University. The team discussed

ohdsi.org/phenotype-phebruary-2023