Meet The Titans

OHDSI Community Call
Nov. 7, 2023 • 11 am ET

Upcoming Community Calls

Date	Topic
Nov. 7	Meet The Titans
Nov. 14	Collaborator Showcase Honorees
Nov. 21	Showcase Software Demos
Nov. 28	TBA
Dec. 5	Recent Publications
Dec. 12	Happy Birthday OHDSI! Where Have We Come In 10 Years, and in 12 Months?
Dec. 19	Holiday-Themed Goodbye to 2023!

Best Community Contribution Honorees!

Data Standards

Augmenting the National COVID Cohort Collaborative (N3C) Dataset with Medicare and Medicaid (CMS) Data, Secure and Deidentified Clinical Dataset

PRESENTER: Stephanie S. Hong

INTRO:

The National COVID Cohort Collaborative (N3C) data Enclave is a platform that provides researchers access to COVID-related patient EMR data in OMOP CDM format. It is the largest centralized repository of COVID-related Patient EMR data in U.S. CMS claims data is also transformed into OMOP CDM format using code map service. N3C COVID patient cohort is now linked to CMS claims data via Privacy Preserving Record Linkage (PPRL). As a result, N3C EMR datasets in OMOP CDM format are enriched with the following additional CMS claims data.

Inpatient	***Outpatient,***
Part D drug prescription	***Home health***
Part B	***Skilled nursing***
Long term care	***Other services***
Durable medical equipment	

METHODS

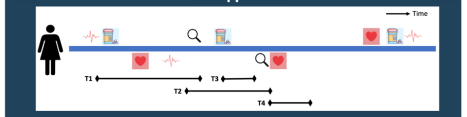
1. CMS claim files in wide format are parsed and pivoted into long format. The clinical concept codes are organized into a condensed format per patient per visit for efficient data transformation.
2. The condensed dataset is then used by the Code Map service to generate the clinical concept translation table. The unified version of the OMOP vocabulary tables are used to perform the translation from the source code to OMOP concept IDs
3. The generated code map service table is used as input in the data pipeline to transform the CMS claims datasets into OMOP CDM format.
4. The data pipeline is built to generate CMS dataset in OMOP CDM format with N3C PPRL linkage.
5. N3C data is enriched with CMS data per PPRL-linked N3C patient. In cases where N3C person_id is duplicated, a Global ID is provided for each.

How much data from Medicare?

How much data from Medicaid?

Link to continuously updated view of N3C enriched dataset via PPRL linkage: https://covid.cd2h.org/dashboard/public-health/pprl/1

A timeline with no gaps, either overlapping or contiguous, is used to construct the *Macro Visit* akin to the N3C macro visit approach.



Take a picture to download the brief report

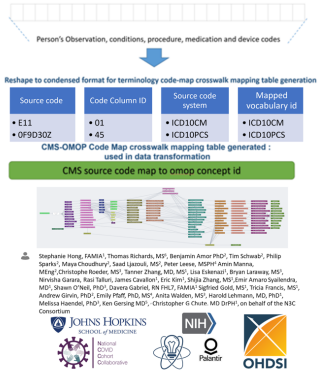
Contact: shong59@jh.edu

RESULT:

82 sites are participating in N3C
30 sites are participating in N3C CMS PPRL linkage
N3C total patients : **20,868,921**
N3C PPRL-linked CMS patients : **359,096**
Total rows of data in N3C: **28.3billion**
Total rows of data in CMS: **653,366,927**
N3C dataset enriched by CMS
Among the PPRL-linked patients, in average, additional concepts are available from CMS:

Claims	Domain	PPRL-Linked patient/ domain	Average # of additional concepts added per patient
Medicare	Condition	71%	78
Medicare	Procedure	60%	6.33
Medicare	Drug_exposure	75%	21.83
Medicare	Measurement	60%	16.48
Medicare	Observation	66.9%	8.6
Medicare	Device	47.6%	6.8
Medicaid	Condition	20.8%	33.9
Medicaid	Procedure	20.2%	23
Medicaid	Drug_exposure	21.9%	20.9
Medicaid	Measurement	17.8%	17.44
Medicaid	Observation	18.3%	6.68
Medicaid	Device	13.9%	6.3

Code Map Service: *Terminology codes appear in multiple columns, i.e.col01 to col45. And some claim source files were over 4000 columns wide. The dataset is pivoted to condense format to generate the clinical concept translation table using OMOP vocab.*



Stephanie Hong, FAMIA[1], Thomas Richards, MS[2], Benjamin Amor PhD[2], Tim Schwab[2], Philip Sparks[2], Maya Choudhury[2], Saad Ljazouli, MS[2], Peter Leese, MSPH[2] Amin Manna, MEng[2],Christophe Roeder, MS[2], Tanner Zhang, MD, MS[2], Lisa Eskenazi[2], Bryan Laraway, MS[2], Nirvisha Garara, Rasi Talluri, James Cavallon[2], Eric Kim[2], Shijia Zhang, MS[2] Emir Amaro Syailendra, MD[2], Shawn O'Neil, PhD[2], Davera Gabriel, RN FHL7, FAMIA[2] Sigfried Gold, MS[2], Tricia Francis, MS[2], Andrew Girvin, PhD[2], Emily Pfaff, PhD, MS[2], Anita Walden, MS[2], Harold Lehmann, MD, PhD[2], Melissa Haendel, PhD[2], Ken Gersing MD[2], Christopher G Chute. MD DrPH[2], on behalf of the N3C Consortium

Augmenting the National COVID Cohort Collaborative (N3C) Dataset with Medicare and Medicaid (CMS) Data, Secure and Deidentified Clinical Dataset

(Stephanie Hong, Thomas Richards, Benjamin Amor, Tim Schwab, Philip Sparks, Maya Choudhury, Saad Ljazouli, Peter Leese, Amin Manna, Christophe Roeder, Tanner Zhang, Lisa Eskenazi, Bryan Laraway, James Cavallon, Eric Kim, Shijia Zhang, Emir Amaro Syailendra, Shawn O'Neil, Davera Gabriel, Sigfried Gold, Tricia Francis, Andrew Girvin, Emily Pfaff, Anita Walden, Harold Lehmann, Melissa Haendel, Ken Gersing, Christopher G Chute)

Best Community Contribution Honorees!

Methods Research

Generating Synthetic Electronic Health Records in OMOP using GPT

Chao Pang[1], Xinzhuo Jiang[1], Nishanth Parameshwar Pavinkurve[1], Krishna S. Kalluri[1], Elise L. Minto[1], Karthik Natarajan[1]
[1]Columbia University Irving Medical Center, Department of Biomedical Informatics

OHDSI

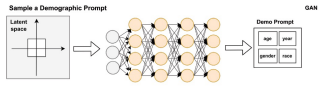
Background

This work focuses on synthetic data generation and demonstrate the capability of training a GPT model using a patient representation derived from CEHR-BERT, enabling the generation of patient sequences that can be seamlessly converted to the OMOP data format bi-direction.

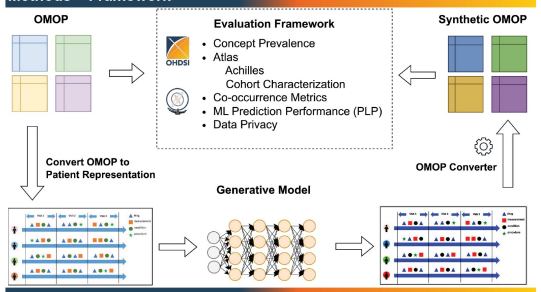
Current approach: Bag of Word + GAN Model

Use cases of synthetic EHR data:

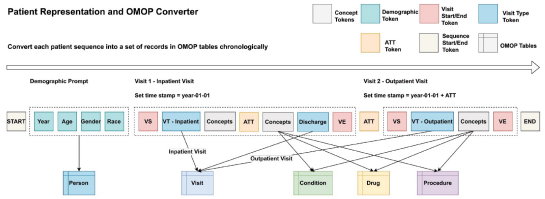
- Phenotype algorithm validation
- Prediction research
- Tool development
- External validation
- Training and education
- Debiasing the the source data
- Counterfactual dataset



Methods – Framework

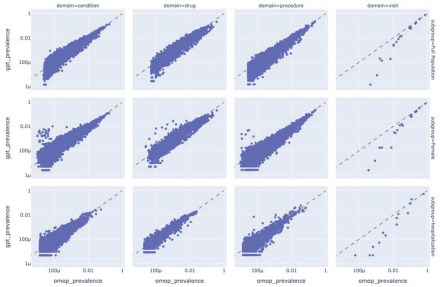


Methods – Patient Representation and OMOP Converter

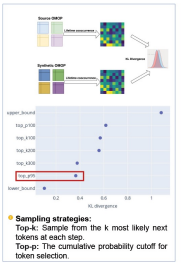


Results

Concept Prevalence



Co-occurrence Metrics



Sampling strategies:
Top-k: Sample from the k most likely next tokens at each step.
Top-p: The cumulative probability cutoff for token selection.

Machine Learning Performance Metrics

Target Cohorts	Real data	Top P cumulative probability cutoff P (%)		Top K K concepts with the highest probabilities	
		Top P = 95%	Top P = 100%	Top K = 100	Top K = 300
HF Readmission	Pre = 25.7 AUC = 65.7 PR = 39.3	Pre = 27.6 AUC = 69.2 PR = 45.7	Pre = 28.4 AUC = 65.9 PR = 41.8	Pre = 30.7 AUC = 68.1 PR = 47.8	Pre = 26.5 AUC = 64.9 PR = 39.3
Hospitalization	Pre = 5.6 AUC = 75.3 PR = 19.5	Pre = 5.2 AUC = 77.1 PR = 21.4	Pre = 7.3 AUC = 68.3 PR = 16.5	Pre = 2.8 AUC = 87.0 PR = 22.1	Pre = 6.3 AUC = 78.7 PR = 24.6
COPD Readmission	Pre = 34.5 AUC = 74.2 PR = 83.8	Pre = 37.8 AUC = 76.4 PR = 84.4	Pre = 47.2 AUC = 74.1 PR = 67.2	Pre = 26.4 AUC = 75.9 PR = 90.3	Pre = 34.5 AUC = 68.8 PR = 80.2
Afib Ischemic Stroke	Pre = 8.7 AUC = 84.0 PR = 48.5	Pre = 10.2 AUC = 78.9 PR = 41.2	Pre = 10.4 AUC = 70.7 PR = 39.1	Pre = 16.6 AUC = 77.1 PR = 50.5	Pre = 10.8 AUC = 76.8 PR = 38.5
CAD CABG	Pre = 7.1 AUC = 88.4 PR = 55.9	Pre = 4.1 AUC = 81.5 PR = 25.2	Pre = 4.4 AUC = 52.9 PR = 4.3	Pre = 7.2 AUC = 75.6 PR = 38.5	Pre = 4.0 AUC = 79.0 PR = 24.1

Conclusions

- First framework generated longitudinal synthetic EHR data using OMOP CDM.
- Designed an innovative patient representation by incorporating temporal information which allowed for an accurate reconstruction of patient medical timeline as compared to state of art methods.
- Comprehensive evaluation procedures showed that the synthetic data preserved the fundamental patient characteristics of the real population.

Contact: CEHR-BERT@lists.cumc.columbia.edu

Generating Synthetic Electronic Health Records in OMOP using GPT

(Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Elise L. Minto, Karthik Natarajan)

Best Community Contribution Honorees!

Open-Source Development

GUSTO Data Vault:
Laying the foundations for an open science system with OMOP Data Catalogue

PRESENTER: Cindy Ho, Mukkesh Kumar

INTRO:
- Growing Up in Singapore Towards healthy Outcomes (GUSTO) aims to understand how conditions in pregnancy and early childhood influence the subsequent health and development of women and children.
- The A*STAR/GUSTO Data Vault platform have advanced data exploration capabilities for research data, biospecimens and publications asset management.
- The OMOP Data Catalogue was created in GUSTO Data Vault to showcase the GUSTO data which have been converted into OMOP CDM format.

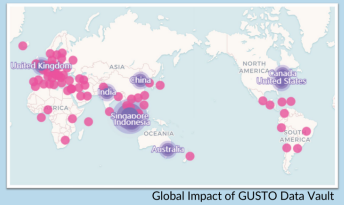
METHODS
- Data Vault (containerized web application with Docker) was built using PostgreSQL database and Django.
- Tools used: HTML, CSS, jQuery, Ajax, Python, Plotly Dash, Dashboard engine in Dash Enterprise, AWS Cloud Platform.
- OMOP fields were mapped using Athena and customized R programming scripts.

RESULTS
- OMOP Data Catalogue makes GUSTO cohort-specific CDM fields to be discovered across the Person, Condition, Observation and Measurement tables by the global research community.
- Metadata is described with relevant attributes such as CDM Field, Concept ID, Name, Subject Type, Visit Timepoint, Description and Domain.
- Data profiling of the OMOP Concept IDs enables GUSTO data to be reused, described, discovered, and identified by researchers (FAIR data principles).
- OMOPed data from incremental OMOP conversions can be seamlessly integrated in OMOP Data Catalogue by GUSTO data curators.
- This enables database level characterizations for GUSTO study.

GUSTO OMOP Data Catalogue lays the foundations for developing cross-study OMOP Data Catalogues expanded across **APAC and global OHDSI data partners**, enabling **database level characterizations.**

Scan to visit GUSTO Data Vault (https://gustodatavault.sg)

Scan to download the abstract



Global Impact of GUSTO Data Vault

Our future work includes the optimization of GUSTO OMOP data conversion journey using advanced OMOP conversion tools such as the IQVIA OMOP Converter.

Snippets of OMOP Data Catalogue

Landing Page

Person table

Measurement table

Observation table

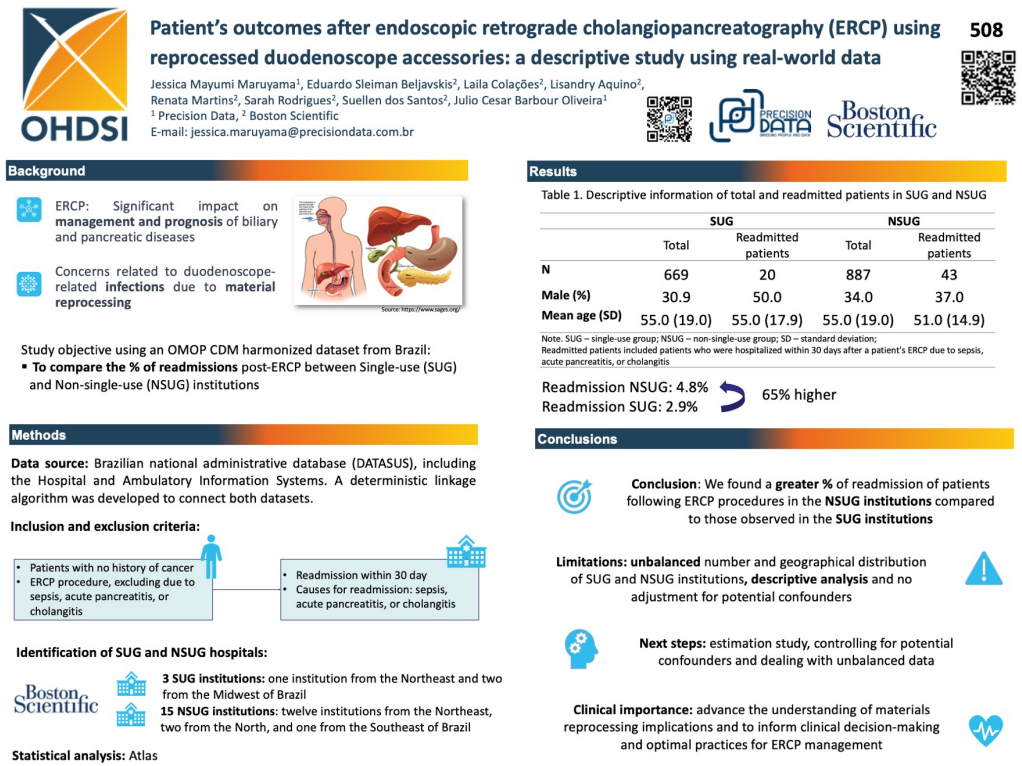
Cindy Ho, Li Ting Ang, Maisie Ng, Hang Png, Shuen Lin Tan, Estella Ye, Sunil Kumar Raja, Mengling Feng, Johan G Eriksson, Mukkesh Kumar

Agency for Science, Technology and Research SINGAPORE
NUS Saw Swee Hock School of Public Health
OHDSI

GUSTO Data Vault: Laying the foundations for an open science system with OMOP Data Catalogue (Cindy Ho, Li Ting Ang, Maisie Ng, Hang Png, Shuen Lin Tan, Estella Ye, Sunil Kumar Raja, Mengling Feng, Johan G Eriksson, Mukkesh Kumar)

Best Community Contribution Honorees!

Clinical Applications



OHDSI

Patient's outcomes after endoscopic retrograde cholangiopancreatography (ERCP) using reprocessed duodenoscope accessories: a descriptive study using real-world data

508

Jessica Mayumi Maruyama[1], Eduardo Sleiman Beljavskis[2], Laila Colações[2], Lisandry Aquino[2], Renata Martins[2], Sarah Rodrigues[2], Suellen dos Santos[2], Julio Cesar Barbour Oliveira[1]
[1] Precision Data, [2] Boston Scientific
E-mail: jessica.maruyama@precisiondata.com.br

PRECISION DATA

Boston Scientific

Background

ERCP: Significant impact on **management and prognosis** of biliary and pancreatic diseases

Concerns related to duodenoscope-related **infections** due to **material reprocessing**

Source: https://www.sages.org/

Study objective using an OMOP CDM harmonized dataset from Brazil:
- **To compare the % of readmissions** post-ERCP between Single-use (SUG) and Non-single-use (NSUG) institutions

Methods

Data source: Brazilian national administrative database (DATASUS), including the Hospital and Ambulatory Information Systems. A deterministic linkage algorithm was developed to connect both datasets.

Inclusion and exclusion criteria:

- Patients with no history of cancer
- ERCP procedure, excluding due to sepsis, acute pancreatitis, or cholangitis

- Readmission within 30 day
- Causes for readmission: sepsis, acute pancreatitis, or cholangitis

Identification of SUG and NSUG hospitals:

Boston Scientific

3 SUG institutions: one institution from the Northeast and two from the Midwest of Brazil
15 NSUG institutions: twelve institutions from the Northeast, two from the North, and one from the Southeast of Brazil

Statistical analysis: Atlas

Results

Table 1. Descriptive information of total and readmitted patients in SUG and NSUG

	SUG		NSUG	
	Total	Readmitted patients	Total	Readmitted patients
N	669	20	887	43
Male (%)	30.9	50.0	34.0	37.0
Mean age (SD)	55.0 (19.0)	55.0 (17.9)	55.0 (19.0)	51.0 (14.9)

Note. SUG – single-use group; NSUG – non-single-use group; SD – standard deviation;
Readmitted patients included patients who were hospitalized within 30 days after a patient's ERCP due to sepsis, acute pancreatitis, or cholangitis

Readmission NSUG: 4.8%
Readmission SUG: 2.9%
65% higher

Conclusions

Conclusion: We found a **greater %** of readmission of patients following ERCP procedures in the **NSUG institutions** compared to those observed in the **SUG institutions**

Limitations: unbalanced number and geographical distribution of SUG and NSUG institutions**, descriptive analysis** and no adjustment for potential confounders

Next steps: estimation study, controlling for potential confounders and dealing with unbalanced data

Clinical importance: advance the understanding of materials reprocessing implications and to inform clinical decision-making and optimal practices for ERCP management

Patient's outcomes after endoscopic retrograde cholangiopancreatography (ERCP) using reprocessed duodenoscope accessories: a descriptive study using real-world data

(Jessica Mayumi Maruyama, Eduardo Sleiman Beljavskis, Laila Colações, Lisandry Aquino, Renata Martins, Sarah Rodrigues, Suellen dos Santos, Julio Cesar Barbour Oliveira)

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	9 am	Patient-Level Prediction
Wednesday	12 pm	Health Equity
Wednesday	2 pm	Natural Language Processing
Thursday	8 am	India Chapter
Thursday	9:30 am	Data Network Quality
Thursday	12 pm	Medical Devices
Thursday	7 pm	Dentistry
Friday	9 am	Phenotype Development & Evaluation
Friday	9 am	GIS – Geographic Information System Development
Friday	1 pm	Clinical Trials
Friday	11 pm	China Chapter
Monday	9 am	Vaccine Vocabulary
Monday	10 am	Africa Chapter
Monday	11 am	Early-Stage Researchers
Tuesday	9 am	OMOP CDM Oncology Genomic Subgroup

Global Symposium Homepage

2023 OHDSI Symposium

Oct. 20-22 · East Brunswick, New Jersey

The 2023 OHDSI Global Symposium welcomed more than 440 of our global collaborators together for three days of sharing research, forging new connections and pushing forward together the OHDSI mission of improving health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care.

This page will be home to all materials from the global symposium. Check back in the coming days for all video presentations from the event! #JoinTheJourney #OHDSI2023




State of the Community

Various leaders within OHDSI shared a presentation on the state of the community, with specific focuses on data standards, vocabulary enhancements and open-source development. **Speakers included**:

George Hripcsak, Columbia University
Clair Blacketer, Johnson & Johnson
Alexander Davydov, Odysseus Data Services
Katy Sadowski, Boehringer Ingelheim
Peter Rijnbeek, Erasmus MC
Mengling 'Mornin' Feng, National University of Singapore

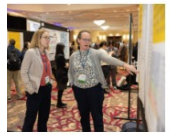
State of the Community Slides

Collaborator Showcase Posters & Software Demos

ceived a record number of submissions for the 2023 Collaborator case, following detailed review by community volunteers in the Scientific w Committee, there were 137 posters and 24 software demos that were nted during the collaborator showcase.

e visit the link below to visit the posters, brief reports and other ementary materials for each showcase submission. Each submission will filed in the #OHDSISocialShowcase, so please make sure you follow l on Twitter/X, LinkedIn and Instagram.

2023 Collaborator Showcase Posters & Demos




Tutorial: Introduction to OHDSI

urney from data to evidence can be challenging alone but is greatly nabled through community collaboration. In this half-day tutorial, we will introduce newcomers to OHDSI. Specifically, about the tools, practices, and open-science approach to evidence generation that the OHDSI community has developed and evolved over the past decade.

Faculty will highlight the ways community individuals can participate as well as receive value from the community's outputs. The course will include topics such as open community data standards – including the OMOP Common Data Model and OHDSI Standardized Vocabularies, opensource analytic tools

2023 Global Collaborator Showcase

Observational Data Standards & Management

2 – FinOMOP – a population-based data network (Javier Gracia-Tabuenca, Perttu Koskenvesa, Pia Tajanen, Sampo Kukkurainen, Gustav Klingstedt, Anna Hammais, Persephone Doupi, Oscar Brück, Leena Hakkarainen, Annu Kaila, Marco Hautalahti, Toni Mikkola, Marianna Niemi, Pasi Rikala, Simo Ryhänen, Anna Virtanen, Arto Mannermaa, Arto Vuori, Joanne Demmler, Eric Fey, Terhi Kilpi, Arho Virkki, Tarja Laitinen, Kimmo Porkka)
3 – From OMOP to CDISC SDTM: Successes, Challenges, and Future Opportunities of using EHR Data for Drug Repurposing in COVID-19 (Wesley Anderson, Ruth Kurtycz, Tahsin Farid, Shermarke Hassan, Kalynn Kennon, Pam Dasher, Danielle Boyce, Will Roddy, Smith F. Heavner)
4 – Augmenting the National COVID Cohort Collaborative (N3C) Dataset with Medicare and Medicaid (CMS) Data, Secure and Deidentified Clinical Dataset (Stephanie Hong, Thomas Richards, Benjamin Amor, Tim Schwab, Philip Sparks, Maya Choudhury, Saad Ljazouli, Peter Leese, Amin Manna, Christophe Roeder, Tanner Zhang, Lisa Eskenazi, Bryan Laraway, James Cavallon, Eric Kim, Shijia Zhang, Emir Amaro Syailendra, Shawn O'Neil, Davera Gabriel, Sigfried Gold, Tricia Francis, Andrew Girvin, Emily Pfaff, Anita Walden, Harold Lehmann, Melissa Haendel, Ken Gersing, Christopher G Chute)
5 – Integrating clinical and laboratory research data using the OMOP CDM (Edward A. Frankenberger, Chun Yang, Vamsidhar Reddy Meda Venkata, Alyssa Goodson)
6 – Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension (Woo Yeon Park, Kyulee Jeon, Teri Sippel Schmidt, Haridimos Kondylakis, Seng Chan You, Paul Nagy)
7 – Conversion of a Myositis Precision Medicine Center into a Common Data Model: A Case Study (Zachary Wang, Will Kelly, Paul Nagy, Christopher A Mecoli)
8 – Implementing a common data model in ophthalmology: Comparison of general eye examination mapping to standard OMOP concepts across two major EHR systems (Justin C. Quon, William Halfpenny, Cindy X. Cai, Sally L. Baxter, Brian C. Toy)
9 – Enhancing Data Quality Management: Introducing Capture and Cleanse Modes to the Data Quality Dashboard (Frank DeFalco, Clair Blacketer)
10 – "OMOP Anywhere": Daily Updates from EHR Data Leveraging Epic's Native Tools (Mujeeb A Basit, Mereeja Varghese, Aamirah Vadsariya, Bhavini Nayee, Margaret Langley, Ashley Huynh, Jennifer Cai, Donglu Xie, Cindy Kao, Eric Nguyen, Todd Boutte, Shiby Antony, Tammye Garrett, Christoph U Lehmann, Duwayne L Willett)
11 – A Toxin Vocabulary for the OMOP CDM (Maksym Trofymenko, Polina Talapova, Tetiana Nesmiian, Andrew Williams, Denys Kaduk, Max Ved, Inna Ageeva)
12 – Challenges and opportunities in adopting OMOP-CDM in Brazilian healthcare: a report from Hospital Israelita Albert Einstein (Maria Abrahao, Uri Adrian Prync Flato, Mateus de Lima Freitas, Diogo Patrão, Amanda Gomes Rabelo, Cesar Augusto Madid Truyts, Gabriela Chiuffa Tunes, Etienne Duin, Gabriel Mesquita de Souza, Soraya Yukari Aashiro, Adriano José Pereira, Edson Amaro)
13 – Transforming the Optum® Enriched Oncology module to OMOP CDM (Dmitry Dymshyts, Clair Blacketer)
14 – Mapping Multi-layered Oncology Data in OMOP (John Methot, Sherry Lee)
15 – Development of psychiatric common data model (P-CDM) leveraging psychiatric scales (Dong Yun Lee, Chungsoo Kim, Rae Woong Park)
16 – Brazilian administrative data for real-world research: a deterministic linkage procedure and OMOP CDM harmonization (Jessica Mayumi Maruyama, Julio Cesar Barbour Oliveira)
17 – Integration of Clinical and Genomic Data Mapped to the OMOP Common Data Model in a Federated Data Network in Belgium (Tatjana Jatsenko, Murat Akand, Joris Robert Vermeesch, Dries Rombaut, Michel Van Speybroeck, Martine Lewi , Valerie Vandeweerd)

ohdsi.org/OHDSI2023

November Newsletter

The Journey Newsletter (November 2023)

The 2023 OHDSI Global Symposium welcomed more than 430 of our global collaborators together for three days of sharing research, forging new connections and pushing forward the OHDSI mission of improving health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care. This newsletter will highlight all the available materials from #OHDSI2023 (including videos, slides & posters), while sharing some of the best moments from our three days together. **#JoinTheJourney**

OHDSI Videocast: 2023 Symposium Review

In the latest On The Journey video, Patrick Ryan and Craig Sachson reflect on the OHDSI 2023 Global Symposium, including looks at the State of the Community talk, and both the plenary (Improving the reliability and scale of case validation) and panel (Lessons learned from OHDSI network studies) sessions. They also discuss the largest collaborator showcase in OHDSI history, the weekend events, and more. *(if video does not appear, click 'view this email in your browser')*

Community Updates

Where Have We Been?

- The 2023 Global Symposium was held Oct. 20-22 in East Brunswick, N.J., and it included a main conference and a full weekend of activities, including the "Welcome to OHDSI" tutorial. All presentations, slides and showcase submissions from OHDSI2023 are now available on the event homepage.
- The 2023 Titan Awards were presented at the OHDSI Global Symposium. Congratulations to our 2023 honorees, who were both nominated and selected by fellow community members!

- Data Standards: **Gowtham Rao** and **Azza Shoaibi**
- Methodological Research: **Jiayi (Jessie) Tong**
- Open-Source Development: **Katy Sadowski**
- Clinical Applications: **Center for Surgical Science**
- Community Leadership: **Nicole Pratt**
- Community Collaboration: **Cynthia Sung**
- Community Support: **Gyeol Song**

Where Are We Now?

- Thirty-six organizations and data sources were introduced as the original members of the OHDSI Evidence Network during the State of the Community presentation. These are among the 534 data sources mapped to the OMOP CDM. If your data source would like to join the OHDSI Evidence Network, please fill out this brief form.
- The latest edition of the Our Journey annual report was shared at the Global Symposium, and there is a digital version that you can access on the OHDSI website.
- The #OHDSISocialShowcase has begun for the Global Symposium. Please make sure you are following OHDSI's LinkedIn, Twitter/X and Instagram feeds to receive daily updates on the research presented by our community.

OHDSI 2023 Focuses On Large-Scale Evidence Generation & Community Collaboration

The 2023 OHDSI Global Symposium brought together more than 430 community members from around the world for a three-day event filled with opportunities to learn, connect and forge new relationships.

The main conference was held during Day 1, and featured a plenary on improving the reliability and scale of case validation, a State of the Community presentation by several leaders in the community, and a panel on lessons learned from OHDSI network studies. The collaborator showcase included a record number of posters, software demos and lightning talks, and the closing included a interactive session on large-scale collaboration, escape-room style.

There were also two days of workshops, workgroup meetings and an Introduction to OHDSI tutorial. Videos and slides from all presentations and the tutorial are available on the symposium homepage. Thank you to those who both volunteered their time to make the event a success or joined us to help push forward OHDSI's mission of improving health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care.

Symposium Homepage

Collaborator Showcase Posters & Software Demos

Spotlight: Atif Adam

While my time in the OHDSI community is relatively brief, the environment struck a chord with me immediately. What stands out is how the community welcomes expertise from myriad backgrounds. Whether you're a seasoned researcher, a data scientist, a clinician, or even someone just starting in healthcare analytics, OHDSI is a platform where different levels of familiarity converge to nurture actionable knowledge.

Dr. Atif Adam is a systems scientist and researcher boasting over a decade of diversified experience spanning academia, industry, and entrepreneurial ventures. He attained his doctorate in Health Systems Science and Computational Epidemiology. In addition, Dr. Adam completed his clinical training in Internal Medicine and secured master's degrees in Health Policy and Spatial Epidemiology.

His research probes the nuanced relationships between chronic cardiometabolic diseases, mental health, cognitive aging, and health disparities. During his academic appointments at institutions such as Johns Hopkins and Harvard, Dr. Adam pioneered innovative simulation frameworks

October Publications

Frid S, Bracons Cucó G, Gil Rojas J, López-Rueda A, Pastor Duran X, Martínez-Sáez O, Lozano-Rubí R. Evaluation of OMOP CDM, i2b2 and ICGC ARGO for supporting data harmonization in a breast cancer use case of a multicentric European AI project. J Biomed Inform. 2023 Sep 27;147:104505. doi: 10.1016/j.jbi.2023.104505. Epub ahead of print. PMID: 37774908.

Kim JW, Kim C, Kim KH, Lee Y, Yu DH, Yun J, Baek H, Park RW, You SC. Scalable Infrastructure Supporting Reproducible Nationwide Healthcare Data Analysis toward FAIR Stewardship. Sci Data. 2023 Oct 4;10(1):674. doi: 10.1038/s41597-023-02580-7. PMID: 37794003; PMCID: PMC10550904.

Oh S, Joo HJ, Sohn JW, Park S, Jang JS, Seong J, Park KJ, Lee SH. Cloud-based digital healthcare development for precision medical hospital information system. Per Med. 2023 Sep;20(5):435-444. doi: 10.2217/pme-2023-0074. Epub 2023 Oct 9. PMID: 37811595.

Sadsad R, Ruber G, Zhou J, Nicklin S, Tsafnat G. A computable biomedical knowledge object for calculating in-hospital mortality for patients admitted with acute myocardial infarction. Learn Health Syst. 2023 Sep 11;7(4):e10388. doi: 10.1002/lrh2.10388. PMID: 37860059; PMCID: PMC10582239.

Krastev E, Tcharaktchiev D, Abanos S. Application of OMOP Common Data Model for Data Integration: The Bulgarian Diabetes Register. Stud Health Technol Inform. 2023 Oct 20;309:141-142. doi: 10.3233/SHTI230761. PMID: 37869827.

Ostropolets A, Hripcsak G, Husain SA, Richter LR, Spotnitz M, Elhussein A, Ryan PB. Scalable and interpretable alternative to chart review for phenotype evaluation using standardized structured data from electronic health records. J Am Med Inform Assoc. 2023 Oct 17:ocad202. doi: 10.1093/jamia/ocad202. Epub ahead of print. PMID: 37847668.

mailchi.mp/ohdsi/november2023

OHDSI at AMIA

Saturday, Nov 11, 2023

8:30 AM - 4:30 PM CST

OHDSI RWE Revolution: Igniting Data Modernization with Harmonized Standards for Cutting-Edge Health Research

In response to the 21st Century Cures Act and the growing importance of Real-world evidence (RWE), the FDA has released guidance on RWE data for regulatory decision-making. Despite RWE's potential to improve clinical studies, challenges remain in rapidly utilizing RWE for decision-making due to the volume and diversity of real-world data, further emphasized by the COVID-19 pandemic. The Observational Health Data Science and Informatics (OHDSI) developed the Observational Medical Outcomes Partnership (OMOP) to address these challenges and ensure the quality of RWE. OMOP focuses on the development of a common data model and standardized analytics to facilitate meaningful comparisons across different RWE data sources and research studies. Building on the need to better understand RWE and OMOP, the workshop gathers leading experts from various fields to discuss three major themes: (1) understanding the origin and barriers of real-world data for healthcare research and the role that OHDSI/OMOP has played in improving the use of RWE for healthcare research; (2) showcasing the potential of RWE analysis across multiple data types with OMOP CDM; and (3) discussing the challenges and opportunities to adopt RWE for secondary use in research and development. Participants will engage with in-depth topics such as data transformation, cohort definitions, diagnostic methods, visualization techniques, and practical applications of cohort diagnostics in real-world scenarios. This event aligns closely with the broader informatics interests of the attendees. It aims to enhance their understanding of the synergies and opportunities at the intersection of OHDSI, OMOP CDM, and healthcare.

Speaker(s):

Speaker: Atif Adam, PhD, MD, MPH, Johns Hopkins Bloomberg School of Public Health Speaker: Mui Van Zandt, MS, IQVIA Speaker: Paul Nagy Speaker: Mengling Feng, PhD, National University of Singapore Speaker: Christian Reich, PhD, MD, OHDSI Speaker: Zhen Lin, PhD, Memorial Herman Texas Medical Center

Workshop - Collaborative

Location: Churchill A

Session Code: W02

Session Credits: 6.00

OHDSI HADES releases: SelfControlledCaseSeries 5.0.0

SelfControlledCaseSeries 5.0.0 | Reference | Articles | Changelog | HADES

SelfControlledCaseSeries

R-CMD-check passing | codecov 87%

SelfControlledCaseSeries is part of HADES.

Introduction

SelfControlledCaseSeries is an R package for performing Self-Controlled Case Series (SCCS) analyses in an observational database in the OMOP Common Data Model.

Features

- Extracts the necessary data from a database in OMOP Common Data Model format.
- Optionally add seasonality using a spline function.
- Optionally add age using a spline function.
- Optionally add calendar time using a spline function.
- Optionally correct for event-dependent censoring of the observation period.
- Optionally add many covariates in one analysis (e.g. all drugs).
- Options for constructing different types of covariates and risk windows, including pre-exposure windows (to capture contra-indications).
- Optionally use regularization on all covariates except the outcome of interest.

Links

Browse source code

Report a bug

Ask a question

License

Apache License 2.0

Citation

Citing SelfControlledCaseSeries

Developers

Martijn Schuemie
Author, maintainer

Patrick Ryan
Author

Trevor Shaddox
Author

Marc Suchard
Author

OHDSI HADES releases: DeepPatientLevelPrediction 2.0.1

DeepPatientLevelPrediction 2.0.1 | Get started | My first deep learning model | Reference | Articles | Changelog | HADES

DeepPatientLevelPrediction

R-CMD-check passing | codecov 99%

Introduction

DeepPatientLevelPrediction is an R package for building and validating deep learning patient-level predictive models using data in the OMOP Common Data Model format and OHDSI PatientLevelPrediction framework.

Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969-975.

Features

- Adds deep learning models to use in the OHDSI PatientLevelPrediction framework.
- Allows to add custom deep learning models.
- Includes an MLP, ResNet and a Transformer
- Allows to use all the features of PatientLevelPrediction to validate and explore your model performance.

Technology

Links

Browse source code

Report a bug

Ask a question

License

Apache License 2.0

Citation

Citing DeepPatientLevelPrediction

Developers

Egill Fridgeirsson
Author, maintainer

Jenna Reps
Author

Seng Chan You
Author

Chungsoo Kim
Author

Henrik John

OHDSI HADES releases: DataQualityDashboard 2.5.0

DataQualityDashboard 2.5.0 | Get started | Reference | Articles | Changelog | HADES

DataQualityDashboard

DataQualityDashboard is part of HADES.

The goal of the Data Quality Dashboard (DQD) project is to design and develop an open-source tool to expose and evaluate observational data quality.

Introduction

This package will run a series of data quality checks against an OMOP CDM instance (currently supports v5.4, v5.3 and v5.2). It systematically runs the checks, evaluates the checks against some pre-specified threshold, and then communicates what was done in a transparent and easily understandable way.

Overview

The quality checks were organized according to the Kahn Framework[1] which uses a system of categories and contexts that represent strategies for assessing data quality. For an introduction to the kahn framework please click here.

Using this framework, the Data Quality Dashboard takes a systematic-based approach to running data quality checks. Instead of writing thousands of individual checks, we use "data quality check types". These "check types" are more general, parameterized data quality checks into which OMOP tables, fields, and concepts can be substituted to represent a singular data quality idea. For example, one check type might be written as

Links

Browse source code

Report a bug

Ask a question

DQD Example Output

License

Apache License (>= 2)

Citation

Citing DataQualityDas

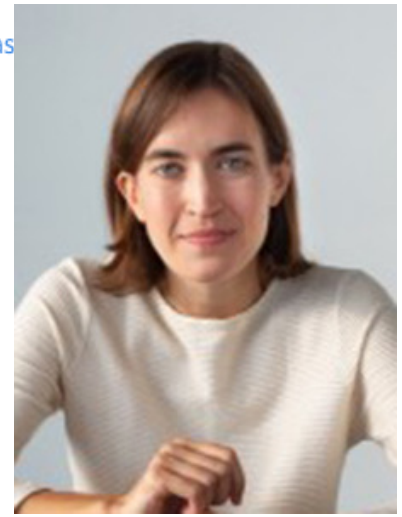
Developers

Katy Sadowski
Author, maintainer

Clair Blacketer
Author

Ajit Londhe
Author

Anthony Sena
Author



OHDSI HADES releases: CohortExplorer 0.1.0

CohortExplorer 0.1.0 | Reference | Articles | Changelog | HADES

CohortExplorer

R-CMD-check passing | codecov 100% | CRAN 0.1.0 | downloads 319/month

CohortExplorer is part of HADES.

Introduction

This software tool is designed to extract data from a randomized subset of individuals within a cohort and make it available for exploration in a 'Shiny' application environment. It retrieves date-stamped, event-level records from one or more data sources that represent patient data in the Observational Medical Outcomes Partnership (OMOP) data model format. This tool features a user-friendly interface that enables users to efficiently explore the extracted profiles, thereby facilitating applications, such as reviewing structured profiles. The output of this R-package is a self-contained R shiny that contains person-level data for review.

Warning

- Contains person level data. This package is not to be considered de-identified.
- Please do not share the output with others as it may violate protected health information.
- .RDS file in output contains PHI.

Links

View on CRAN

Browse source code

Report a bug

Ask a question

License

Apache License

Citation

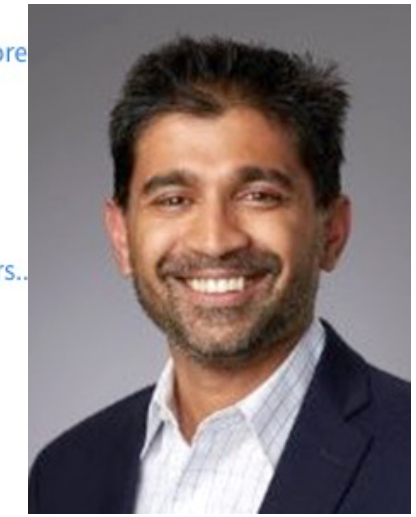
Citing CohortExplore

Developers

Gowtham Rao

Author, maintainer

More about authors..



#OHDSISocialShowcase This Week

MONDAY

Conversion of a Myositis Precision Medicine Center into a Common Data Model: A Case Study

(Zachary Wang, Will Kelly, Paul Nagy, Christopher A Mecoli)

Conversion of a Myositis Precision Medicine Center into a Common Data Model: *A Case Study*

INTRO:

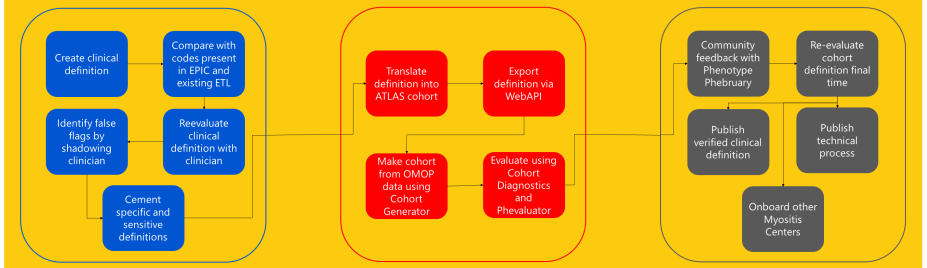
- Myositis is a rare disease and most centers lack the numbers to answer questions on their own. **Enter: OMOP!**
- The problem is that most guidance for "OMOP-fying" a dataset are not geared towards small physician-led tertiary care centers working inside of a larger institution or in highly constrained IT environments.
- Our goals were to 1) Add our center's registry data to the JHU ETL. 2) Develop phenotypes to isolate our cohorts of interest. 3) Blaze a path for other Myositis centers to act as research and data partners.

METHODS AND RESULTS

After encountering numerous obstacles we developed a pipeline for ETL'ing our data with these generalizable steps:

1. Assemble the Team; identify and bring in subject matter specialists.
2. Figure out how many of our concepts can be represented in the CDM vocab using USAGI.
3. Spend a LOT of time mapping and validating concepts with USAGI and Athena.
4. Rally-the-Troops in Myositis centers around the world to agree on custom concepts not available in the standard vocabulary (still ongoing).
5. Get CohortDiagnostics to actually run in our highly constrained computing environment.
6. Work with our institution to make the process easier for the next team.

"OMOP-ifying" data from inside an academic institution is hard... Here's how we did it!



Take a picture to download the full abstract

Additional information:

- Some of the issues we had to work around:
  - WebAuth not compatible with Vegas authentication used in our Atlas instance.
    - Copy/Pasted to/from the OHDSI atlas demo instance.
  - DatabaseConnector not compatible with our ActiveDirectory setup.
    - Used a portable R installation inside our Windows Environment to generate the SQLite files to feed into CohortExplorer.
  - The environment we were required to use had an older version of R and no easy way to update it.
    - Worked with IT to have R upgraded.
  - Windows auth not supported in Linux.
    - Developed custom code to override the DatabaseConnector package.
  - Lots of Java workarounds.
    - Workarounds with IT to find compatible Java setup.
    - "Black magic" code from the OHDSI forums.
  - Lack of technical specialties in the center.
    - Recruiting grad students to assist with specialized tasks.
    - Connected with institutional OMOP team.
    - Training lead physician investigator to work with OHDSI tools.
    - Attended formal OMOP course.
  - No funding for project.
    - Looking for grants continue work.

Vocabulary mapping:

- Time spent: 45hrs by staff and 5hrs by faculty.
- 325 concepts mapped.
- 125 custom concepts.

Scannable Links:

Authors:

Zachary Wang MS, Will Kelly, Paul Nagy PhD, Christopher A Mecoli MD, MHS

JOHNS HOPKINS UNIVERSITY

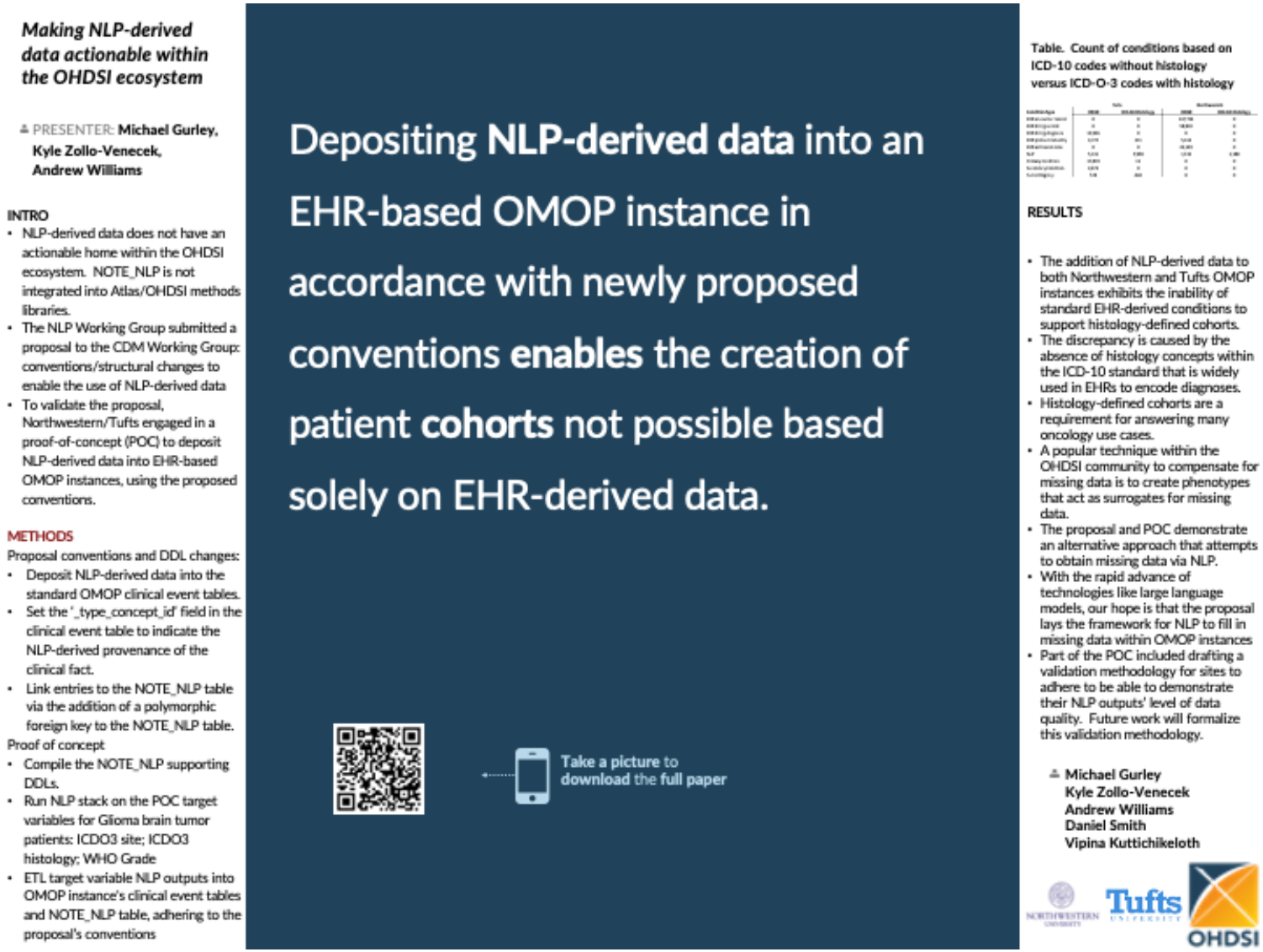
OHDSI

#OHDSISocialShowcase This Week

TUESDAY

Making NLP-derived data actionable within the OHDSI ecosystem

(Michael Gurley, Kyle Zollo-Venecek, Andrew Williams, Daniel Smith, Robert Miller, Vipina Kuttichi Keloth, Hua Xu)



Making NLP-derived data actionable within the OHDSI ecosystem

PRESENTER: **Michael Gurley, Kyle Zollo-Venecek, Andrew Williams**

INTRO

- NLP-derived data does not have an actionable home within the OHDSI ecosystem. NOTE_NLP is not integrated into Atlas/OHDSI methods libraries.
- The NLP Working Group submitted a proposal to the CDM Working Group: conventions/structural changes to enable the use of NLP-derived data
- To validate the proposal, Northwestern/Tufts engaged in a proof-of-concept (POC) to deposit NLP-derived data into EHR-based OMOP instances, using the proposed conventions.

METHODS

Proposal conventions and DDL changes:

- Deposit NLP-derived data into the standard OMOP clinical event tables.
- Set the '_type_concept_id' field in the clinical event table to indicate the NLP-derived provenance of the clinical fact.
- Link entries to the NOTE_NLP table via the addition of a polymorphic foreign key to the NOTE_NLP table.

Proof of concept

- Compile the NOTE_NLP supporting DDLs.
- Run NLP stack on the POC target variables for Glioma brain tumor patients: ICDO3 site; ICDO3 histology; WHO Grade
- ETL target variable NLP outputs into OMOP instance's clinical event tables and NOTE_NLP table, adhering to the proposal's conventions

Depositing **NLP-derived data** into an EHR-based OMOP instance in accordance with newly proposed conventions **enables** the creation of patient **cohorts** not possible based solely on EHR-derived data.

Take a picture to download the full paper

Table. Count of conditions based on ICD-10 codes without histology versus ICD-O-3 codes with histology

RESULTS

- The addition of NLP-derived data to both Northwestern and Tufts OMOP instances exhibits the inability of standard EHR-derived conditions to support histology-defined cohorts.
- The discrepancy is caused by the absence of histology concepts within the ICD-10 standard that is widely used in EHRs to encode diagnoses.
- Histology-defined cohorts are a requirement for answering many oncology use cases.
- A popular technique within the OHDSI community to compensate for missing data is to create phenotypes that act as surrogates for missing data.
- The proposal and POC demonstrate an alternative approach that attempts to obtain missing data via NLP.
- With the rapid advance of technologies like large language models, our hope is that the proposal lays the framework for NLP to fill in missing data within OMOP instances
- Part of the POC included drafting a validation methodology for sites to adhere to be able to demonstrate their NLP outputs' level of data quality. Future work will formalize this validation methodology.

Michael Gurley
Kyle Zollo-Venecek
Andrew Williams
Daniel Smith
Vipina Kuttichikeloth

Northwestern University · Tufts University · OHDSI

#OHDSISocialShowcase This Week

WEDNESDAY

Demonstrating Utility of the Edge Tool Suite through Clinical Trial Emulation

(Ruth Kurtycz, Wesley Anderson, Allan J. Walkey, Kerry A. Howard, Smith F. Heavner)

Title: Demonstrating Utility of the Edge Tool Suite through Clinical Trial Emulation

PRESENTER: **Ruth Kurtycz**

INTRO:

- Real-world data (RWD) can support repurposing approved medications for new uses. OMOP standardizes RWD, aiding research and sharing, but implementing OMOP can be resource-intensive. The Edge Tool suite streamlines data conversion. Here we assesses the utility of data converted with the Edge Tool suite's potential for research using an emulation of the RECOVERY COVID-19 clinical trial.

METHODS

1. The Edge Tool Suite enables EHR ETL into the OMOP CDM. Propensity score matching at a 2:1 ratio was used to match patients with and without dexamethasone administration on 11 key variables (e.g., patient age)
2. A pilot healthcare site provided over 10,000 acute COVID-19 patient records from March 2020 to March 2022, with exclusion criteria applied to align with the RECOVERY trial.
3. Logistic regression and survival analyses were preformed on the matched data to assess the impact of Dexamethasone on 28-day mortality.

RESULTS

Results	Overall	Oxygen Support	
RECOVERY	Lower 28-day mortality in treatment group	Invasive Mechanical Ventilation: Lower	
		Oxygen Alone: Lower	
		No Oxygen: Not significantly different	
Edge Tool Replication	Higher 28-day mortality in treatment group	Invasive Mechanical Ventilation vs No Oxygen: Lower	Oxygen Alone vs No Oxygen: Lower

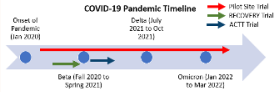
- The analysis found that dexamethasone reduced 28-day mortality in COVID-19 patients receiving oxygen alone or mechanical ventilation, aligning with the RECOVERY trial. However, without oxygen support stratification, this result was not confirmed.

We demonstrate the utility of RWD from the Edge Tool Suite to replicate findings of a clinical trial for COVID-19. Results indicate approach has potential to be used to assess the efficacy of treatments for emerging diseases

Take a picture to download the full paper

AMMO BAR

- Resource-intensive OMOP implementation can exclude smaller healthcare sites, especially in disadvantaged areas. The Edge Tool suite reduces implementation time and costs, making data conversion to OMOP more accessible.
- Below visualization to show overlap of RECOVERY trial and pilot site data



- Exclusion criteria applied to the pilot site records included patients under 18, and pregnant or breastfeeding women. Data quality assessment was also conducted, evaluating missingness, plausibility, and outliers in laboratory values.
- Eleven categorical variables were used in the matching process, followed by evaluation using Chi-Square tests to ensure appropriate balance between treatment and control groups.
- Analyses were performed using binomial logistic regression and survival analysis with a Cox proportional hazards model, repeated after stratifying cases and controls based on the level of oxygen support.
- Challenges in the data extraction process, particularly regarding medication dosage and timing information, may have affected the analysis results.
- Despite limitations, this analysis supports emphasizing the ongoing effort to standardize data using OMOP
- Ongoing updates to the analysis are expected with contributions from more healthcare sites.

Ruth Kurtycz, Wesley Anderson, Allan J. Walkey, Kerry A. Howard, Smith F. Heavner

OHDSI

#OHDSISocialShowcase This Week

THURSDAY

OHDSI Network Study Execution Framework and Templating

(Ben Martin, Cindy Cai, Asieh Golozar, Paul Nagy)

OHDSI Network Study Execution Framework and Templating

PRESENTER: **Ben Martin**

INTRO:

- To enhance robustness and reliability of OHDSI network studies, it is essential to define/outline key steps and status indicators of study development and execution.

OBJECTIVES

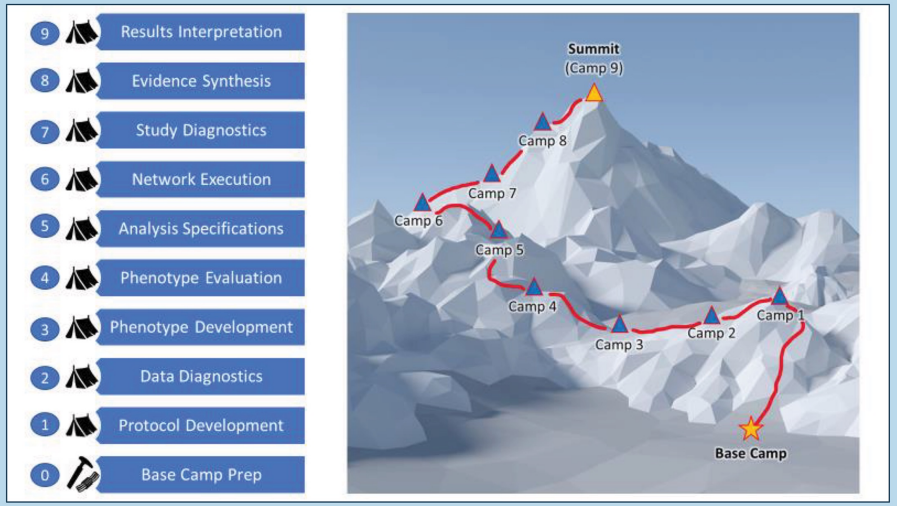
- Define the key process steps common amongst network studies, to improve repeatability and reproducibility and help set expectations for researchers looking to engage in network studies.
- Develop standardized human and computer readable indicators of network study status, progress, and needs.

METHODS

- Consensus from individuals with experience in leading OHDSI network studies demarcated nine fundamental stages that all network studies must progress through towards completion.
- A standard set of human readable and computable data artifacts attributed to each network study are derived from the distillation of these stages and from existing documentation of completed network studies.

The **nine stages of a network study** are illustrated as successive **"camps"** along each **network study "expedition"** in Figure 1, below.

Figure 1. Nine Stages of an OHDSI Network Study Expedition



Take a picture to download the full paper

Table 1. Data Artifacts for Network Study Monitoring

Study Attribute	Values
IRB materials are sufficient for review	[No, Yes]
Cohort definition available	[No, Yes]
Data partner recruitment status	[Not Ready, Open, Closed]
Deadline for adding new data partners	MM/DD/YYY
Statistician partner recruitment status	[Not Ready, Open, Closed]
Deadline for adding new statistician partners	MM/DD/YYY
Clinical domain partner recruitment status	[Not Ready, Open, Closed]
Deadline for adding new clinical domain partners	MM/DD/YYY

RESULTS

- **Figure 1:** nine stages of a network study were identified as, in order: protocol development, data diagnostics, phenotype development, phenotype evaluation, analysis specifications, network execution, study diagnostics, evidence synthesis, results evaluation.
- **Table 1:** a proposed set of data artifacts for study progress monitoring and facilitation.

CONCLUSION

OHDSI network studies share a great degree of common methodology and challenges. Laying out the key steps with a common framework, providing clarity and direction through each of these common stages, and identifying key information for monitoring progress amongst the community will facilitate progression and use shared experience to overcome repetitive challenges. The framework of network study stages and set of study progress artifacts proposed here needs to be refined and internalized by the OHDSI community at large.

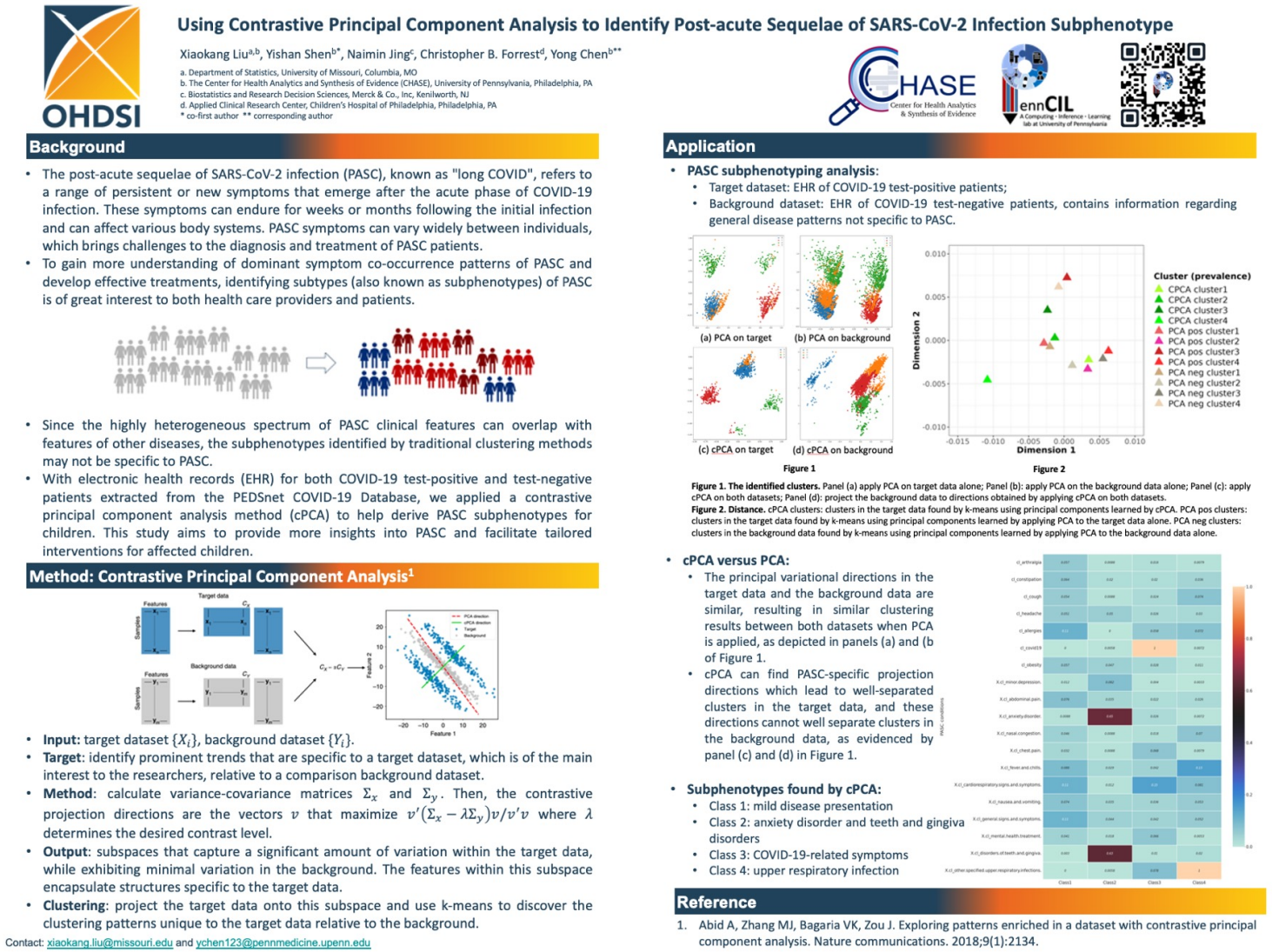
[1]Ben Martin, PhD; [1]Cindy Cai, MD; [2,3]Asieh Golozar, MD; [1]Paul Nagy, PhD;
[1]Johns Hopkins School of Medicine, Baltimore, MD, USA;
[2]Odysseus Data Services, MA, USA;
[3]OHDSI Center at the Roux Institute, Northeastern University, Boston, MA, USA.

#OHDSISocialShowcase This Week

FRIDAY

Using Contrastive Principal Component Analysis to Identify Post-acute Sequelae of SARS-CoV-2 Infection Subphenotypes: an EHR-Based Cohort from the RECOVER Program

(Xiaokang Liu, Yishan Shen, Naimin Jing, Christopher B. Forrest, Yong Chen)



Using Contrastive Principal Component Analysis to Identify Post-acute Sequelae of SARS-CoV-2 Infection Subphenotype

Xiaokang Liu[a,b], Yishan Shen[b*], Naimin Jing[c], Christopher B. Forrest[d], Yong Chen[b**]

a. Department of Statistics, University of Missouri, Columbia, MO
b. The Center for Health Analytics and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA
c. Biostatistics and Research Decision Sciences, Merck & Co., Inc, Kenilworth, NJ
d. Applied Clinical Research Center, Children's Hospital of Philadelphia, Philadelphia, PA
* co-first author ** corresponding author

Background

- The post-acute sequelae of SARS-CoV-2 infection (PASC), known as "long COVID", refers to a range of persistent or new symptoms that emerge after the acute phase of COVID-19 infection. These symptoms can endure for weeks or months following the initial infection and can affect various body systems. PASC symptoms can vary widely between individuals, which brings challenges to the diagnosis and treatment of PASC patients.
- To gain more understanding of dominant symptom co-occurrence patterns of PASC and develop effective treatments, identifying subtypes (also known as subphenotypes) of PASC is of great interest to both health care providers and patients.

- Since the highly heterogeneous spectrum of PASC clinical features can overlap with features of other diseases, the subphenotypes identified by traditional clustering methods may not be specific to PASC.
- With electronic health records (EHR) for both COVID-19 test-positive and test-negative patients extracted from the PEDSnet COVID-19 Database, we applied a contrastive principal component analysis method (cPCA) to help derive PASC subphenotypes for children. This study aims to provide more insights into PASC and facilitate tailored interventions for affected children.

Method: Contrastive Principal Component Analysis[1]

- **Input:** target dataset $\{X_i\}$, background dataset $\{Y_i\}$.
- **Target**: identify prominent trends that are specific to a target dataset, which is of the main interest to the researchers, relative to a comparison background dataset.
- **Method**: calculate variance-covariance matrices Σ_x and Σ_y. Then, the contrastive projection directions are the vectors v that maximize $v'(\Sigma_x - \lambda\Sigma_y)v/v'v$ where λ determines the desired contrast level.
- **Output**: subspaces that capture a significant amount of variation within the target data, while exhibiting minimal variation in the background. The features within this subspace encapsulate structures specific to the target data.
- **Clustering**: project the target data onto this subspace and use k-means to discover the clustering patterns unique to the target data relative to the background.

Contact: xiaokang.liu@missouri.edu and ychen123@pennmedicine.upenn.edu

Application

- **PASC subphenotyping analysis**:
  - Target dataset: EHR of COVID-19 test-positive patients;
  - Background dataset: EHR of COVID-19 test-negative patients, contains information regarding general disease patterns not specific to PASC.

Figure 1 Figure 2

Figure 1. The identified clusters. Panel (a) apply PCA on target data alone; Panel (b): apply PCA on the background data alone; Panel (c): apply cPCA on both datasets; Panel (d): project the background data to directions obtained by applying cPCA on both datasets.
Figure 2. Distance. cPCA clusters: clusters in the target data found by k-means using principal components learned by cPCA. PCA pos clusters: clusters in the target data found by k-means using principal components learned by applying PCA to the target data alone. PCA neg clusters: clusters in the background data found by k-means using principal components learned by applying PCA to the background data alone.

- **cPCA versus PCA:**
  - The principal variational directions in the target data and the background data are similar, resulting in similar clustering results between both datasets when PCA is applied, as depicted in panels (a) and (b of Figure 1.
  - cPCA can find PASC-specific projection directions which lead to well-separated clusters in the target data, and these directions cannot well separate clusters in the background data, as evidenced by panel (c) and (d) in Figure 1.

- **Subphenotypes found by cPCA:**
  - Class 1: mild disease presentation
  - Class 2: anxiety disorder and teeth and gingiva disorders
  - Class 3: COVID-19-related symptoms
  - Class 4: upper respiratory infection

Reference

1. Abid A, Zhang MJ, Bagaria VK, Zou J. Exploring patterns enriched in a dataset with contrastive principal component analysis. Nature communications. 2018;9(1):2134.

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?

Meet The Titans