OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Open Network Studies

OHDSI Community Call
May 21, 2024 • 11 am ET

Upcoming Community Calls

Date	Topic
May 21	Open Studies in the OHDSI Community
May 28	Collaborator Showcase Brainstorm
June 4	NO CALL – EUROPEAN SYMPOSIUM
June 11	European Symposium Review
June 18	Application of LLMs In Evidence Generation Process
June 25	Recent OHDSI Publications

Three Stages of The Journey

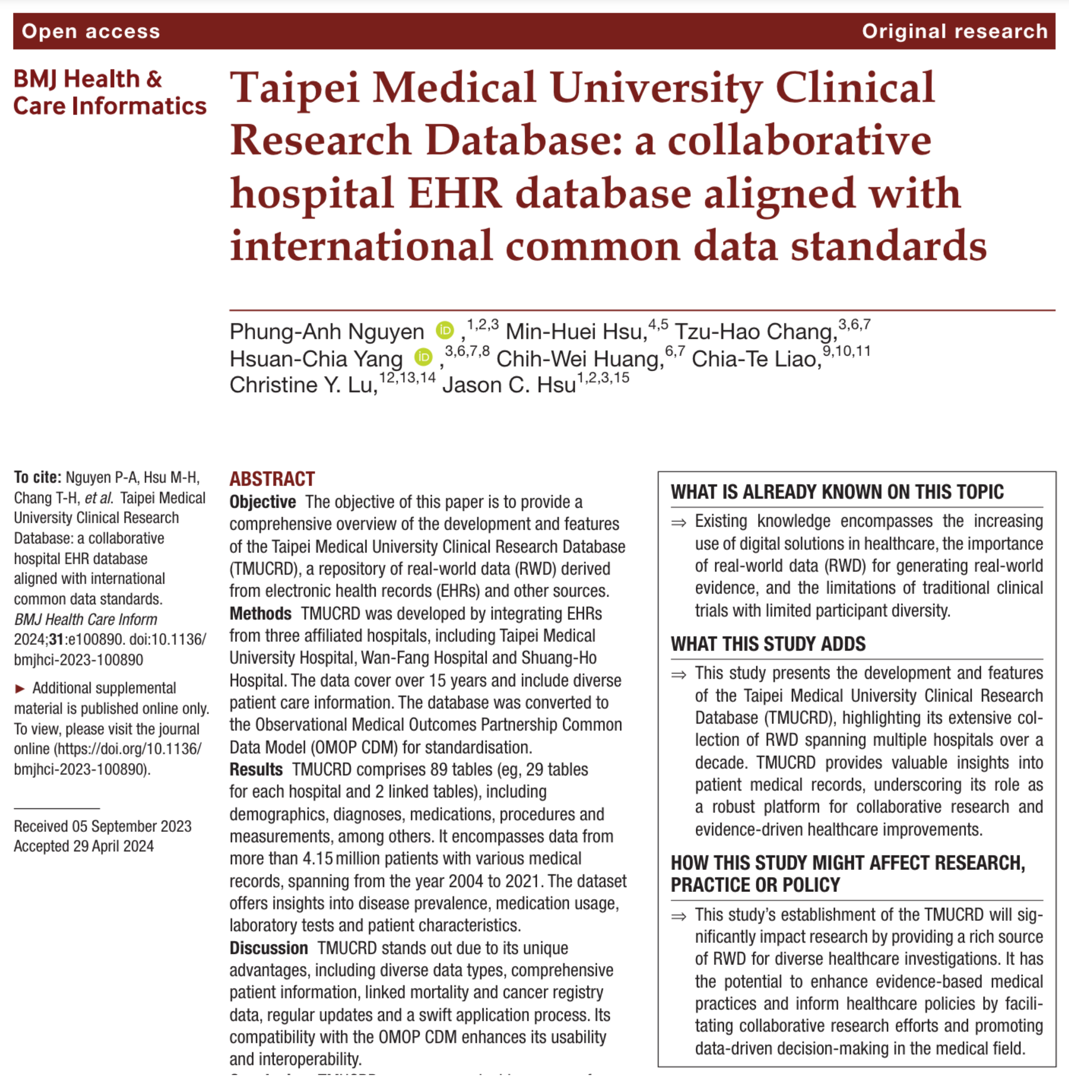
Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Phung-Anh Nguyen, Min-Huei Hsu, Tzu-Hao Chang, Hsuan-Chia Yang, Chih-Wei Huang, Chia-Te Liao, Christine Y. Lu, and Jason C. Hsu** on the publication of **Taipei Medical University Clinical Research Database: a collaborative hospital EHR database aligned with international common data standards** in *BMJ Health & Care Informatics.*



Open access Original research

BMJ Health & Care Informatics

Taipei Medical University Clinical Research Database: a collaborative hospital EHR database aligned with international common data standards

Phung-Anh Nguyen,[1,2,3] Min-Huei Hsu,[4,5] Tzu-Hao Chang,[3,6,7] Hsuan-Chia Yang,[3,6,7,8] Chih-Wei Huang,[6,7] Chia-Te Liao,[9,10,11] Christine Y. Lu,[12,13,14] Jason C. Hsu[1,2,3,15]

To cite: Nguyen P-A, Hsu M-H, Chang T-H, *et al.* Taipei Medical University Clinical Research Database: a collaborative hospital EHR database aligned with international common data standards. *BMJ Health Care Inform* 2024;**31**:e100890. doi:10.1136/bmjhci-2023-100890

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjhci-2023-100890).

Received 05 September 2023
Accepted 29 April 2024

ABSTRACT

Objective The objective of this paper is to provide a comprehensive overview of the development and features of the Taipei Medical University Clinical Research Database (TMUCRD), a repository of real-world data (RWD) derived from electronic health records (EHRs) and other sources.
Methods TMUCRD was developed by integrating EHRs from three affiliated hospitals, including Taipei Medical University Hospital, Wan-Fang Hospital and Shuang-Ho Hospital. The data cover over 15 years and include diverse patient care information. The database was converted to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) for standardisation.
Results TMUCRD comprises 89 tables (eg, 29 tables for each hospital and 2 linked tables), including demographics, diagnoses, medications, procedures and measurements, among others. It encompasses data from more than 4.15 million patients with various medical records, spanning from the year 2004 to 2021. The dataset offers insights into disease prevalence, medication usage, laboratory tests and patient characteristics.
Discussion TMUCRD stands out due to its unique advantages, including diverse data types, comprehensive patient information, linked mortality and cancer registry data, regular updates and a swift application process. Its compatibility with the OMOP CDM enhances its usability and interoperability.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Existing knowledge encompasses the increasing use of digital solutions in healthcare, the importance of real-world data (RWD) for generating real-world evidence, and the limitations of traditional clinical trials with limited participant diversity.

WHAT THIS STUDY ADDS

⇒ This study presents the development and features of the Taipei Medical University Clinical Research Database (TMUCRD), highlighting its extensive collection of RWD spanning multiple hospitals over a decade. TMUCRD provides valuable insights into patient medical records, underscoring its role as a robust platform for collaborative research and evidence-driven healthcare improvements.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study's establishment of the TMUCRD will significantly impact research by providing a rich source of RWD for diverse healthcare investigations. It has the potential to enhance evidence-based medical practices and inform healthcare policies by facilitating collaborative research efforts and promoting data-driven decision-making in the medical field.

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	OMOP CDM Oncology Outreach/Research Subgroup
Tuesday	12 pm	Latin America
Wednesday	3 pm	Joint Vulcan/OHDSI Meeting
Thursday	9:30 am	Network Data Quality
Thursday	7 pm	Dentistry
Friday	9 am	Phenotype Development and Evaluation
Friday	10 am	GIS-Geographic Information System
Friday	11:30 am	Clinical Trials
Friday	11:30 am	Steering Group
Monday	10 am	Africa Chapter
Monday	4 pm	Eyecare & Vision Research
Tuesday	9 am	OMOP CDM Oncology Genomic Subgroup

Next CBER Best Seminar: Tomorrow!

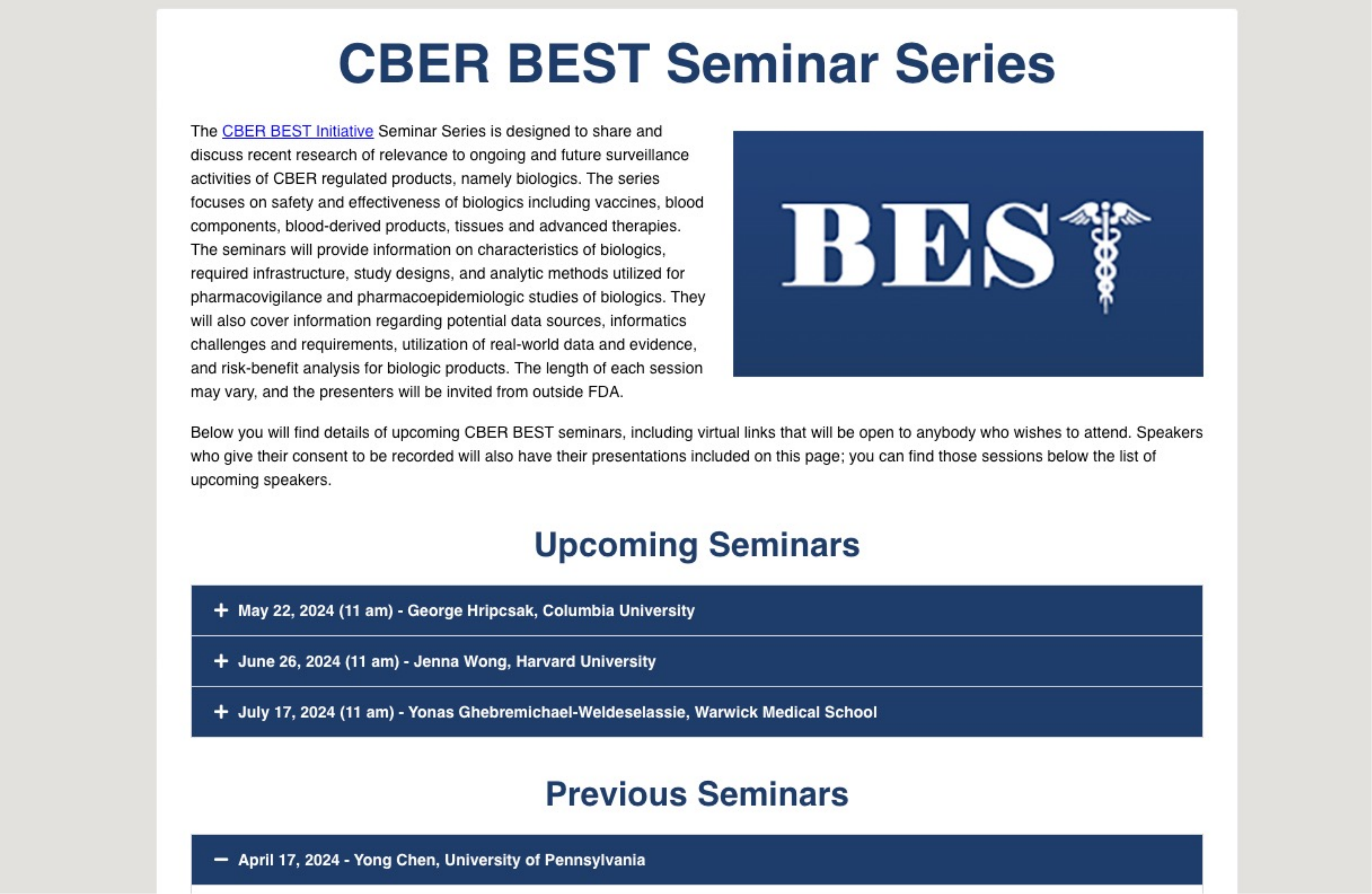
Topic: Reliability in Observational Research: Assessing Covariate Imbalance in Small Studies

Presenter: George Hripcsak, Vivian Beaumont Allen Professor of Biomedical Informatics, Columbia University

Logistics: 11 am – 12 pm EST, Zoom webinar

ohdsi.org/cber-best-seminar-series

CBER Best Seminar Homepage



CBER BEST Seminar Series

The CBER BEST Initiative Seminar Series is designed to share and discuss recent research of relevance to ongoing and future surveillance activities of CBER regulated products, namely biologics. The series focuses on safety and effectiveness of biologics including vaccines, blood components, blood-derived products, tissues and advanced therapies. The seminars will provide information on characteristics of biologics, required infrastructure, study designs, and analytic methods utilized for pharmacovigilance and pharmacoepidemiologic studies of biologics. They will also cover information regarding potential data sources, informatics challenges and requirements, utilization of real-world data and evidence, and risk-benefit analysis for biologic products. The length of each session may vary, and the presenters will be invited from outside FDA.

Below you will find details of upcoming CBER BEST seminars, including virtual links that will be open to anybody who wishes to attend. Speakers who give their consent to be recorded will also have their presentations included on this page; you can find those sessions below the list of upcoming speakers.

Upcoming Seminars

- + May 22, 2024 (11 am) - George Hripcsak, Columbia University
- + June 26, 2024 (11 am) - Jenna Wong, Harvard University
- + July 17, 2024 (11 am) - Yonas Ghebremichael-Weldeselassie, Warwick Medical School

Previous Seminars

- − April 17, 2024 - Yong Chen, University of Pennsylvania

ohdsi.org/cber-best-seminar-series

Kheiron Cohort Application Is Open

The Kheiron Cohort, now in its third year, is a program designed to onboard new contributors into OHDSI and empower them to become active contributors and maintainers.

<u>Career Development</u>

- training opportunities within the cohort from OHDSI technical leaders
- interaction and mentoring from OHDSI leadership

Applications are due June 1

Maternal Health Data Science Fellowship

This program is designed to empower clinical investigators to leverage emerging technologies for improved maternal and neonatal care while reducing morbidity and mortality.

Three main components of this program:

1) Career Development (create evidence, leverage data models, build skills on network studies)

2) Practice (design effective observational research protocols, master tools, write papers/grants)

3) Networking (build relationships with mentors, learners, coordinate with global OHDSI collaborators)

Application deadline: May 22

Want to build your career?

Generate reproducible evidence by leading multi-institutional studies!

Learn more & apply!

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

Peter Robinson, MD

Alexander von Humboldt Professor for AI

Berlin Institute of Health @ Charité

‘The GA4GH Phenopacket Schema: A Standard for Computable Case Reports to Support Translational Genomic Research and Clinical Decision Support Software’

May 30, 2024, 11am-12pm EST

Virtually via Zoom

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

TuftsMedicine
Tufts Medical Center

RWE Workshop at AIME24: Call for Submissions!

Workshop: **AI for Reliable and Equitable Real-World Evidence Generation in Medicine**

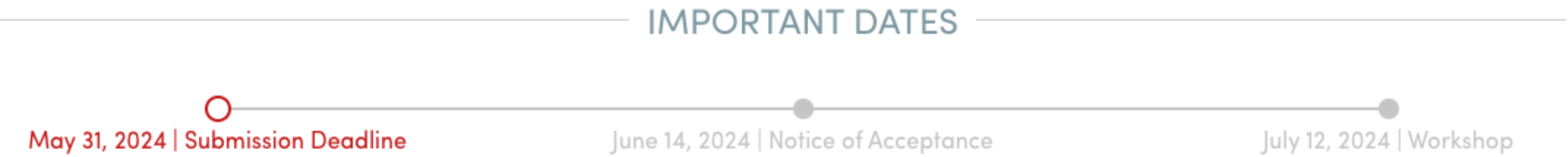
https://medicine.utah.edu/dbmi/aime/ai-reliable

Organizing Committee
Linying Zhang
Adam Wilcox
Yves Lussier

Scientific Program Committee

Peter Rijnbeek	Mattia Prosperi
Larry Han	Xia Ning
Xiaoqian Jiang	Yifan Peng

Opening Keynote
George Hripcsak



IMPORTANT DATES

May 31, 2024 | Submission Deadline

June 14, 2024 | Notice of Acceptance

July 12, 2024 | Workshop

AIME 2024

22nd International Conference on Artificial Intelligence in Medicine
Salt Lake City, Utah, USA, July 9-12

Hosted by the University of Utah

OHDSI Europe Symposium

Registration is OPEN for the **2024 OHDSI Europe Symposium**, which will be held June 1-3 in Rotterdam, Netherlands.

June 1 – tutorial/workshop

June 2 – tutorial/workshop

June 3 – main conference

ohdsi-europe.org

#OHDSI2024 Registration Is Open!

Registration is now OPEN for the 2024 OHDSI Global Symposium, which will be held Oct. 22-24 at the Hyatt Regency Hotel in New Brunswick, N.J., USA.

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

ohdsi.org/OHDSI2024

#OHDSI2024 Collaborator Showcase

Submissions are now being accepted for the 2024 Global Symposium Collaborator Showcase.

All submissions are due by 8 pm ET on Friday, June 21.

Notification of acceptance will be made by Tuesday, Aug. 20.

ohdsi.org/OHDSI2024

#OHDSISocialShowcase This Week

MONDAY

Sirius tool: Conversion of clinical study data into OMOP model and implementation of data quality monitoring of wearable sensor data

(Vojtech Huser, Esteve Verdura, Michael Lubke, Bhavna Adhin)

Sirius tool: Conversion of clinical study data into OMOP model and implementation of data quality monitoring of wearable sensor data

Vojtech Huser MD, PhD, Esteve Verdura MS, Michael N. Lubke, MS, Bhavna Adhin, MS
Pfizer, Inc

BACKGROUND

Optimal data representation of human clinical study data is an ongoing challenge. The Observational Medical Outcomes Partnership (OMOP) model has been used to aggregate data across multiple studies to facilitate analysis that is portable across various datasets.[1] Assessment of data quality of clinical study data, similar to final data analysis, can also be done against OMOP transformed data. Our project focuses on digital health studies that utilize wearable sensors. Digital health technologies significance has been growing recently.[2] Data for wearable sensors is often received and organized into files per subject per study event. The goal of data quality assessment is to look at data file presence (all files present for all study events for all data types for all study participants) and at data file content (files adhere to set of rules that investigate data format, data density, feasibility or context consistency).

METHODS

The data quality assessment framework uses Python and is called ***Sirius***. The name was chosen because the Sirius star, while being a very bright star, also can appear to be changing colors (this results from refraction, which splits the starlight into the colors of a rainbow). We thought that this was similar to the color coding a rule result (e.g., green for compliant, red for errors found).

Sirius uses a modular function approach and this library of functions is extensible to cover different wearable sensor devices and data file formats (see **Table 1**). Sirius data quality rules are defined on study level using Yet Another Markup Language (YAML) syntax (see **Figure 1**). Execution can be set up to be automated for different time intervals (e.g., daily, weekly or monthly execution) and results can be aggregated into a single dashboard view.

RESULTS

Phase 1 of Sirius development took approximately 10 months using a set of six studies that contained wearable sensor data. In phase 2, the library of functions was expanded and the Sirius tool was then applied in 16 additional studies. Sirius either evaluates file presence rules or file content rules. It also uses three types of config files:

1) Study configuration defines study-level metadata. For example, number of study subjects, storage locations to be monitored, or list of expected data sources.
2) Pre-processing actions configuration defines what data transformation should be applied to individual data sources (see **Figure 1**).
3) Rule configuration defines individual rules that evaluate to true (compliant) or false (data error or warning or notification). Actions and rules rely of an extensible set of modular functions. Multiple actions can be chained together to achieve in several steps the necessary data transformation (output of one action becomes input for subsequent action; final action provides input for a data quality rule).

SIRIUS RULE SUPPORTING FUNCTIONS

- **File Name Parsing:** Sirius creates observation events based on parsing the file names that contain the sensor data. This function converts unstructured set of files into database of events assigned to participant and linked to timestamps (OMOP observation table events).
  - Consecutive Visits: For studies where consecutive numbering of visits is used (e.g., visit1 instead of absolute date), it assigns symbolic dates to each visit such that it can be represented in the OMOP model.
  - High Data Frequency: For large sensor data with high frequency of data (more than one data event per minute or hour), the individual rows within sensor file are not converted into formal OMOP events. Subsequent data quality rules then use this OMOP event data to evaluate presence of data per study protocol. An example of a rule is: five cough recording files are present per each visit per each subject.
- **Temporal Data Compliance:** Sirius can analyze temporal patterns in data to detect periods of time when expected sensor data were not recorded (e.g., participant did not wear the sensor [for devices that pause recording during non-wear] or sensor battery was exhausted). The same function also supports detection of outlier values in sensor measurements using multiple outlier identification approaches.
- **Device-specific custom format transformation:** Although most sensors provide directly computable spreadsheet-like data output (e.g., *.CSV or *.H5 parquet format), for sensors using non-standard output, Sirius function library includes pre-processing action functions that facilitate data conversion or data extraction. For example: it can support reading .bin format of ActiGraph device data.

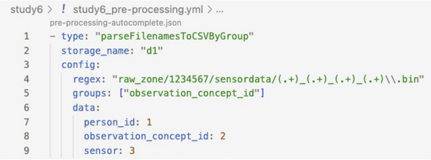
CONCLUSION

We developed a data quality framework for wearable sensor data that automates and improves data monitoring tasks. We also demonstrate that event-based OMOP common data model can facilitate data quality rule authoring for clinical study data.

REFERENCES

1. Roeder C, Sadowski K, Solovyev P, Araujo S. Clinical Trial Data Conventions for the OMOP CDM. In: *OHDSI Symposium*. ; 2020. Available at: https://www.ohdsi.org/2020-global-symposium-showcase-5

2. FDA. Framework for the Use of Digital Health Technologies in Drug and Biological Product Development. Accessed May 16, 2023. Available at: https://fda.gov/digitalhealth


study6 > ! study6_pre-processing.yml > ...
pre-processing-autocomplete.json

```
- type: "parseFilenamesToCSVByGroup"
  storage_name: "d1"
  config:
    regex: "raw_zone/1234567/sensordata/(.+)_(.+)_(.+)_(.+)\\.bin"
    groups: ["observation_concept_id"]
    data:
      person_id: 1
      observation_concept_id: 2
      sensor: 3
```

Figure 1: Authoring a pre-processing action in Visual Studio Code using the function parseFileNamesToCSVByGroup. It employs regular expression and user defines extraction of file name fragments. Autocomplete and error highlighting is achieved using YAML extension.

Function Name	Count
distinctValuesInColumn	34
parseFilenamesToCSV	32
parseFilenamesToCSVByGroup	32
listAsCSV	20
temporalDataComplianceAction	10
containsColumn	7
checkColumnValue	7
checkFileNameFragmentPresenceByVisit	4
countValuesColumnInstancesByGroup	4
countValueColumnInstances	3
combineCSV	3
CSVGroupBy	3
countDistinctValuesPerColumn	2
checkVoltageRange	2
temporalAggregationCSV	2
combineJSONToCSV	1
parseBinFilesToCSV	1
countValuesColumnInstances	1
combineCSVSingle	1
countFilesPerSubject	1
distinctValuesInColumnGroupBy	1
distinctValuesPerFile	1

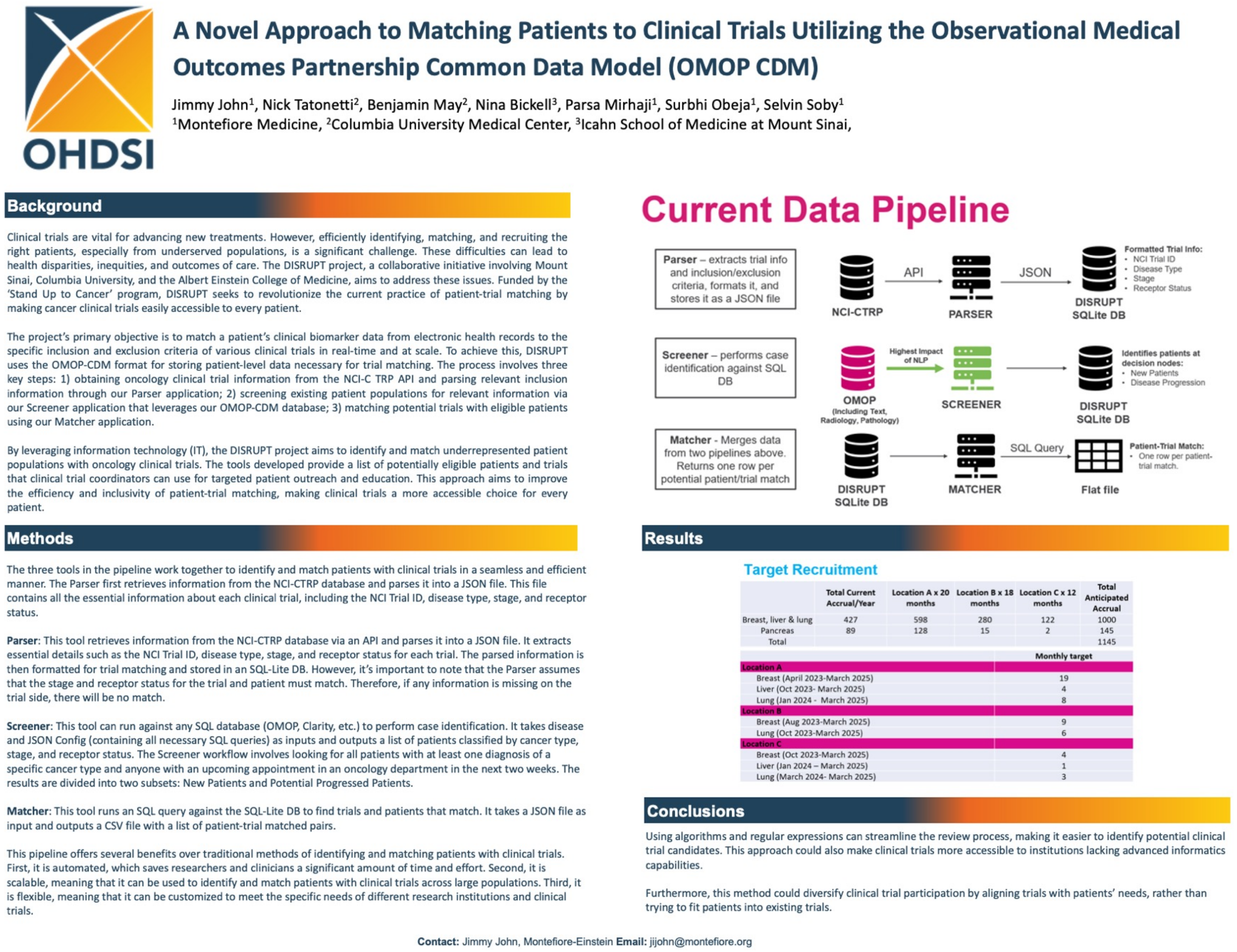
Table 1: List of Sirius Function as well as the frequency of use of those functions across studies (this shows their relative importance).

#OHDSISocialShowcase This Week

TUESDAY

A Novel Approach to Matching Patients to Clinical Trials Using the OMOP Common Data Model

(Jimmy John, Parsa Mirhaji, Surbhi Obeja, Boudewijn Aasman, Nina Bickell, Bruce Rapkin, Erin M. Henninger, Pavel Goriacko, Selvin Soby)



OHDSI

A Novel Approach to Matching Patients to Clinical Trials Utilizing the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)

Jimmy John[1], Nick Tatonetti[2], Benjamin May[2], Nina Bickell[3], Parsa Mirhaji[1], Surbhi Obeja[1], Selvin Soby[1]
[1]Montefiore Medicine, [2]Columbia University Medical Center, [3]Icahn School of Medicine at Mount Sinai,

Background

Clinical trials are vital for advancing new treatments. However, efficiently identifying, matching, and recruiting the right patients, especially from underserved populations, is a significant challenge. These difficulties can lead to health disparities, inequities, and outcomes of care. The DISRUPT project, a collaborative initiative involving Mount Sinai, Columbia University, and the Albert Einstein College of Medicine, aims to address these issues. Funded by the 'Stand Up to Cancer' program, DISRUPT seeks to revolutionize the current practice of patient-trial matching by making cancer clinical trials easily accessible to every patient.

The project's primary objective is to match a patient's clinical biomarker data from electronic health records to the specific inclusion and exclusion criteria of various clinical trials in real-time and at scale. To achieve this, DISRUPT uses the OMOP-CDM format for storing patient-level data necessary for trial matching. The process involves three key steps: 1) obtaining oncology clinical trial information from the NCI-C TRP API and parsing relevant inclusion information through our Parser application; 2) screening existing patient populations for relevant information via our Screener application that leverages our OMOP-CDM database; 3) matching potential trials with eligible patients using our Matcher application.

By leveraging information technology (IT), the DISRUPT project aims to identify and match underrepresented patient populations with oncology clinical trials. The tools developed provide a list of potentially eligible patients and trials that clinical trial coordinators can use for targeted patient outreach and education. This approach aims to improve the efficiency and inclusivity of patient-trial matching, making clinical trials a more accessible choice for every patient.

Methods

The three tools in the pipeline work together to identify and match patients with clinical trials in a seamless and efficient manner. The Parser first retrieves information from the NCI-CTRP database and parses it into a JSON file. This file contains all the essential information about each clinical trial, including the NCI Trial ID, disease type, stage, and receptor status.

Parser: This tool retrieves information from the NCI-CTRP database via an API and parses it into a JSON file. It extracts essential details such as the NCI Trial ID, disease type, stage, and receptor status for each trial. The parsed information is then formatted for trial matching and stored in an SQL-Lite DB. However, it's important to note that the Parser assumes that the stage and receptor status for the trial and patient must match. Therefore, if any information is missing on the trial side, there will be no match.

Screener: This tool can run against any SQL database (OMOP, Clarity, etc.) to perform case identification. It takes disease and JSON Config (containing all necessary SQL queries) as inputs and outputs a list of patients classified by cancer type, stage, and receptor status. The Screener workflow involves looking for all patients with at least one diagnosis of a specific cancer type and anyone with an upcoming appointment in an oncology department in the next two weeks. The results are divided into two subsets: New Patients and Potential Progressed Patients.

Matcher: This tool runs an SQL query against the SQL-Lite DB to find trials and patients that match. It takes a JSON file as input and outputs a CSV file with a list of patient-trial matched pairs.

This pipeline offers several benefits over traditional methods of identifying and matching patients with clinical trials. First, it is automated, which saves researchers and clinicians a significant amount of time and effort. Second, it is scalable, meaning that it can be used to identify and match patients with clinical trials across large populations. Third, it is flexible, meaning that it can be customized to meet the specific needs of different research institutions and clinical trials.

Current Data Pipeline

Parser – extracts trial info and inclusion/exclusion criteria, formats it, and stores it as a JSON file

NCI-CTRP → API → PARSER → JSON → DISRUPT SQLite DB

Formatted Trial Info:
- NCI Trial ID
- Disease Type
- Stage
- Receptor Status

Screener – performs case identification against SQL DB

OMOP (Including Text, Radiology, Pathology) → Highest Impact of NLP → SCREENER → DISRUPT SQLite DB

Identifies patients at decision nodes:
- New Patients
- Disease Progression

Matcher - Merges data from two pipelines above. Returns one row per potential patient/trial match

DISRUPT SQLite DB → MATCHER → SQL Query → Flat file

Patient-Trial Match:
- One row per patient-trial match.

Results

Target Recruitment

	Total Current Accrual/Year	Location A x 20 months	Location B x 18 months	Location C x 12 months	Total Anticipated Accrual
Breast, liver & lung	427	598	280	122	1000
Pancreas	89	128	15	2	145
Total					1145

	Monthly target
Location A	
Breast (April 2023-March 2025)	19
Liver (Oct 2023- March 2025)	4
Lung (Jan 2024 - March 2025)	8
Location B	
Breast (Aug 2023-March 2025)	9
Lung (Oct 2023-March 2025)	6
Location C	
Breast (Oct 2023-March 2025)	4
Liver (Jan 2024 – March 2025)	1
Lung (March 2024- March 2025)	3

Conclusions

Using algorithms and regular expressions can streamline the review process, making it easier to identify potential clinical trial candidates. This approach could also make clinical trials more accessible to institutions lacking advanced informatics capabilities.

Furthermore, this method could diversify clinical trial participation by aligning trials with patients' needs, rather than trying to fit patients into existing trials.

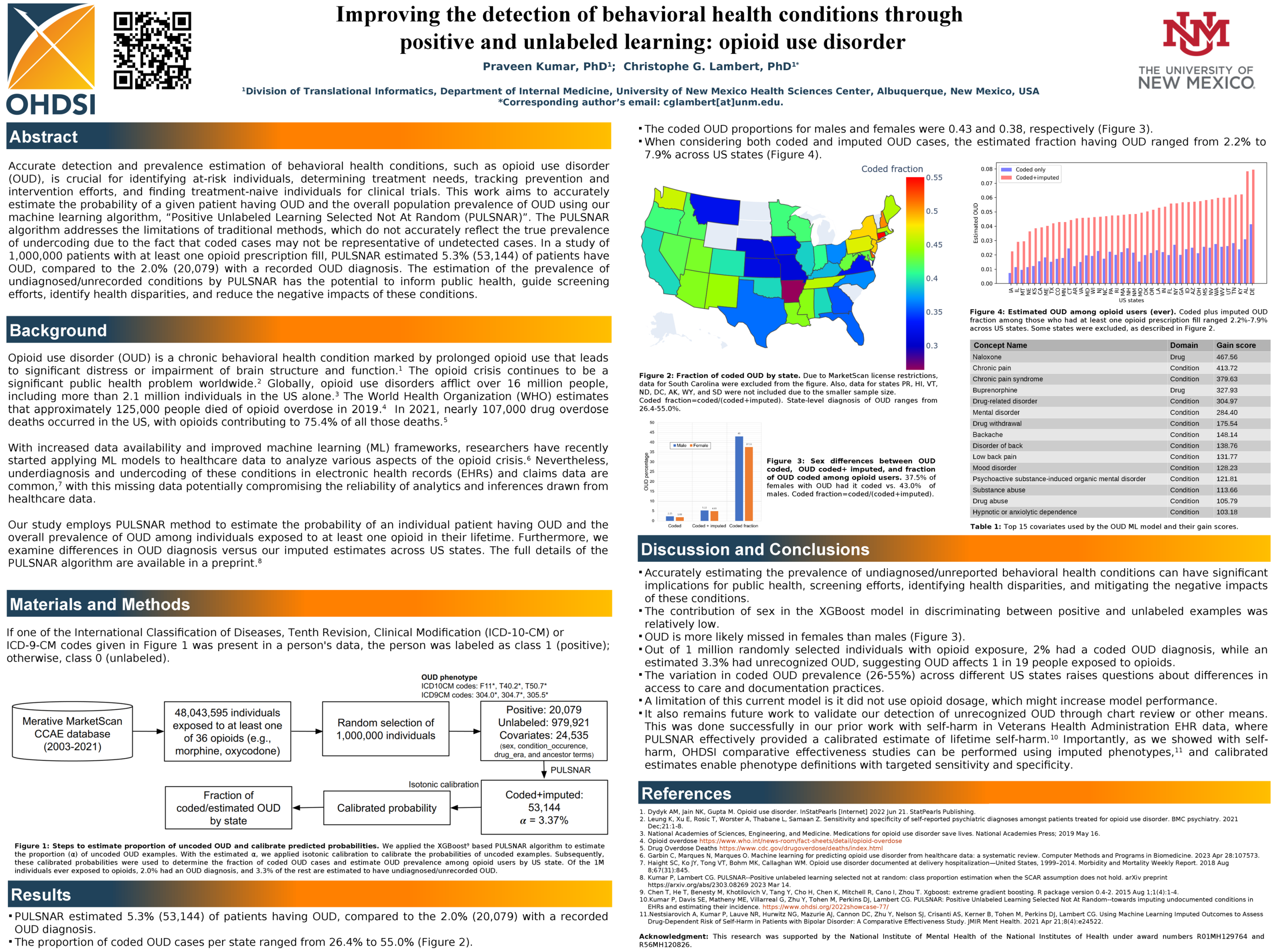
Contact: Jimmy John, Montefiore-Einstein **Email:** jijohn@montefiore.org

#OHDSISocialShowcase This Week

WEDNESDAY

Improving the detection of behavioral health conditions through positive and unlabeled learning: opioid use disorder

(Praveen Kumar, Christophe G. Lambert)



Improving the detection of behavioral health conditions through positive and unlabeled learning: opioid use disorder

Praveen Kumar, PhD[1]; Christophe G. Lambert, PhD[1*]

[1]Division of Translational Informatics, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, New Mexico, USA
*Corresponding author's email: cglambert[at]unm.edu.

Abstract

Accurate detection and prevalence estimation of behavioral health conditions, such as opioid use disorder (OUD), is crucial for identifying at-risk individuals, determining treatment needs, tracking prevention and intervention efforts, and finding treatment-naive individuals for clinical trials. This work aims to accurately estimate the probability of a given patient having OUD and the overall population prevalence of OUD using our machine learning algorithm, "Positive Unlabeled Learning Selected Not At Random (PULSNAR)". The PULSNAR algorithm addresses the limitations of traditional methods, which do not accurately reflect the true prevalence of undercoding due to the fact that coded cases may not be representative of undetected cases. In a study of 1,000,000 patients with at least one opioid prescription fill, PULSNAR estimated 5.3% (53,144) of patients have OUD, compared to the 2.0% (20,079) with a recorded OUD diagnosis. The estimation of the prevalence of undiagnosed/unrecorded conditions by PULSNAR has the potential to inform public health, guide screening efforts, identify health disparities, and reduce the negative impacts of these conditions.

Background

Opioid use disorder (OUD) is a chronic behavioral health condition marked by prolonged opioid use that leads to significant distress or impairment of brain structure and function.[1] The opioid crisis continues to be a significant public health problem worldwide.[2] Globally, opioid use disorders afflict over 16 million people, including more than 2.1 million individuals in the US alone.[3] The World Health Organization (WHO) estimates that approximately 125,000 people died of opioid overdose in 2019.[4] In 2021, nearly 107,000 drug overdose deaths occurred in the US, with opioids contributing to 75.4% of all those deaths.[5]

With increased data availability and improved machine learning (ML) frameworks, researchers have recently started applying ML models to healthcare data to analyze various aspects of the opioid crisis.[6] Nevertheless, underdiagnosis and undercoding of these conditions in electronic health records (EHRs) and claims data are common,[7] with this missing data potentially compromising the reliability of analytics and inferences drawn from healthcare data.

Our study employs PULSNAR method to estimate the probability of an individual patient having OUD and the overall prevalence of OUD among individuals exposed to at least one opioid in their lifetime. Furthermore, we examine differences in OUD diagnosis versus our imputed estimates across US states. The full details of the PULSNAR algorithm are available in a preprint.[8]

Materials and Methods

If one of the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) or ICD-9-CM codes given in Figure 1 was present in a person's data, the person was labeled as class 1 (positive); otherwise, class 0 (unlabeled).

OUD phenotype
ICD10CM codes: F11*, T40.2*, T50.7*
ICD9CM codes: 304.0*, 304.7*, 305.5*

Merative MarketScan CCAE database (2003-2021)

48,043,595 individuals exposed to at least one of 36 opioids (e.g., morphine, oxycodone)

Random selection of 1,000,000 individuals

Positive: 20,079
Unlabeled: 979,921
Covariates: 24,535
(sex, condition_occurence, drug_era, and ancestor terms)

PULSNAR

Coded+imputed: 53,144
α = 3.37%

Isotonic calibration

Calibrated probability

Fraction of coded/estimated OUD by state

Figure 1: Steps to estimate proportion of uncoded OUD and calibrate predicted probabilities. We applied the XGBoost[9] based PULSNAR algorithm to estimate the proportion (α) of uncoded OUD examples. With the estimated α, we applied isotonic calibration to calibrate the probabilities of uncoded examples. Subsequently, these calibrated probabilities were used to determine the fraction of coded OUD cases and estimate OUD prevalence among opioid users by US state. Of the 1M individuals ever exposed to opioids, 2.0% had an OUD diagnosis, and 3.3% of the rest are estimated to have undiagnosed/unrecorded OUD.

Results

- PULSNAR estimated 5.3% (53,144) of patients having OUD, compared to the 2.0% (20,079) with a recorded OUD diagnosis.
- The proportion of coded OUD cases per state ranged from 26.4% to 55.0% (Figure 2).
- The coded OUD proportions for males and females were 0.43 and 0.38, respectively (Figure 3).
- When considering both coded and imputed OUD cases, the estimated fraction having OUD ranged from 2.2% to 7.9% across US states (Figure 4).

Figure 2: Fraction of coded OUD by state. Due to MarketScan license restrictions, data for South Carolina were excluded from the figure. Also, data for states PR, HI, VT, ND, DC, AK, WY, and SD were not included due to the smaller sample size.
Coded fraction=coded/(coded+imputed). State-level diagnosis of OUD ranges from 26.4-55.0%.

Figure 3: Sex differences between OUD coded, OUD coded+ imputed, and fraction of OUD coded among opioid users. 37.5% of females with OUD had it coded vs. 43.0% of males. Coded fraction=coded/(coded+imputed).

Coded only
Coded+imputed
Estimated OUD
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
IA IL MT NE KS CA ME TX CO MN CT AR VA MD WI NJ NC PA RI MA NH NM MO OK OR LA IN FL NY GA ID AZ OH MS NV WA WV UT TN KY AL DE
US states

Figure 4: Estimated OUD among opioid users (ever). Coded plus imputed OUD fraction among those who had at least one opioid prescription fill ranged 2.2%-7.9% across US states. Some states were excluded, as described in Figure 2.

Concept Name	Domain	Gain score
Naloxone	Drug	467.56
Chronic pain	Condition	413.72
Chronic pain syndrome	Condition	379.63
Buprenorphine	Drug	327.93
Drug-related disorder	Condition	304.97
Mental disorder	Condition	284.40
Drug withdrawal	Condition	175.54
Backache	Condition	148.14
Disorder of back	Condition	138.76
Low back pain	Condition	131.77
Mood disorder	Condition	128.23
Psychoactive substance-induced organic mental disorder	Condition	121.81
Substance abuse	Condition	113.66
Drug abuse	Condition	105.79
Hypnotic or anxiolytic dependence	Condition	103.18

Table 1: Top 15 covariates used by the OUD ML model and their gain scores.

Discussion and Conclusions

- Accurately estimating the prevalence of undiagnosed/unreported behavioral health conditions can have significant implications for public health, screening efforts, identifying health disparities, and mitigating the negative impacts of these conditions.
- The contribution of sex in the XGBoost model in discriminating between positive and unlabeled examples was relatively low.
- OUD is more likely missed in females than males (Figure 3).
- Out of 1 million randomly selected individuals with opioid exposure, 2% had a coded OUD diagnosis, while an estimated 3.3% had unrecognized OUD, suggesting OUD affects 1 in 19 people exposed to opioids.
- The variation in coded OUD prevalence (26-55%) across different US states raises questions about differences in access to care and documentation practices.
- A limitation of this current model is it did not use opioid dosage, which might increase model performance.
- It also remains future work to validate our detection of unrecognized OUD through chart review or other means. This was done successfully in our prior work with self-harm in Veterans Health Administration EHR data, where PULSNAR effectively provided a calibrated estimate of lifetime self-harm.[10] Importantly, as we showed with self-harm, OHDSI comparative effectiveness studies can be performed using imputed phenotypes,[11] and calibrated estimates enable phenotype definitions with targeted sensitivity and specificity.

References

1. Dydyk AM, Jain NK, Gupta M. Opioid use disorder. InStatPearls [Internet] 2022 Jun 21. StatPearls Publishing.
2. Leung K, Xu E, Rosic T, Worster A, Thabane L, Samaan Z. Sensitivity and specificity of self-reported psychiatric diagnoses amongst patients treated for opioid use disorder. BMC psychiatry. 2021 Dec;21:1-8.
3. National Academies of Sciences, Engineering, and Medicine. Medications for opioid use disorder save lives. National Academies Press; 2019 May 16.
4. Opioid overdose https://www.who.int/news-room/fact-sheets/detail/opioid-overdose
5. Drug Overdose Deaths https://www.cdc.gov/drugoverdose/deaths/index.html
6. Garbin C, Marques N, Marques O. Machine learning for predicting opioid use disorder from healthcare data: a systematic review. Computer Methods and Programs in Biomedicine. 2023 Apr 28:107573.
7. Haight SC, Ko JY, Tong VT, Bohm MK, Callaghan WM. Opioid use disorder documented at delivery hospitalization—United States, 1999–2014. Morbidity and Mortality Weekly Report. 2018 Aug 8;67(31):845.
8. Kumar P, Lambert CG. PULSNAR--Positive unlabeled learning selected not at random: class proportion estimation when the SCAR assumption does not hold. arXiv preprint https://arxiv.org/abs/2303.08269 2023 Mar 14.
9. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T. Xgboost: extreme gradient boosting. R package version 0.4-2. 2015 Aug 1;1(4):1-4.
10. Kumar P, Davis SE, Matheny ME, Villarreal G, Zhu Y, Tohen M, Perkins DJ, Lambert CG. PULSNAR: Positive Unlabeled Learning Selected Not At Random--towards imputing undocumented conditions in EHRs and estimating their incidence. https://www.ohdsi.org/2022showcase-77/
11. Nestsiarovich A, Kumar P, Lauve NR, Hurwitz NG, Mazurie AJ, Cannon DC, Zhu Y, Nelson SJ, Crisanti AS, Kerner B, Tohen M, Perkins DJ, Lambert CG. Using Machine Learning Imputed Outcomes to Assess Drug-Dependent Risk of Self-Harm in Patients with Bipolar Disorder: A Comparative Effectiveness Study. JMIR Ment Health. 2021 Apr 21;8(4):e24522.

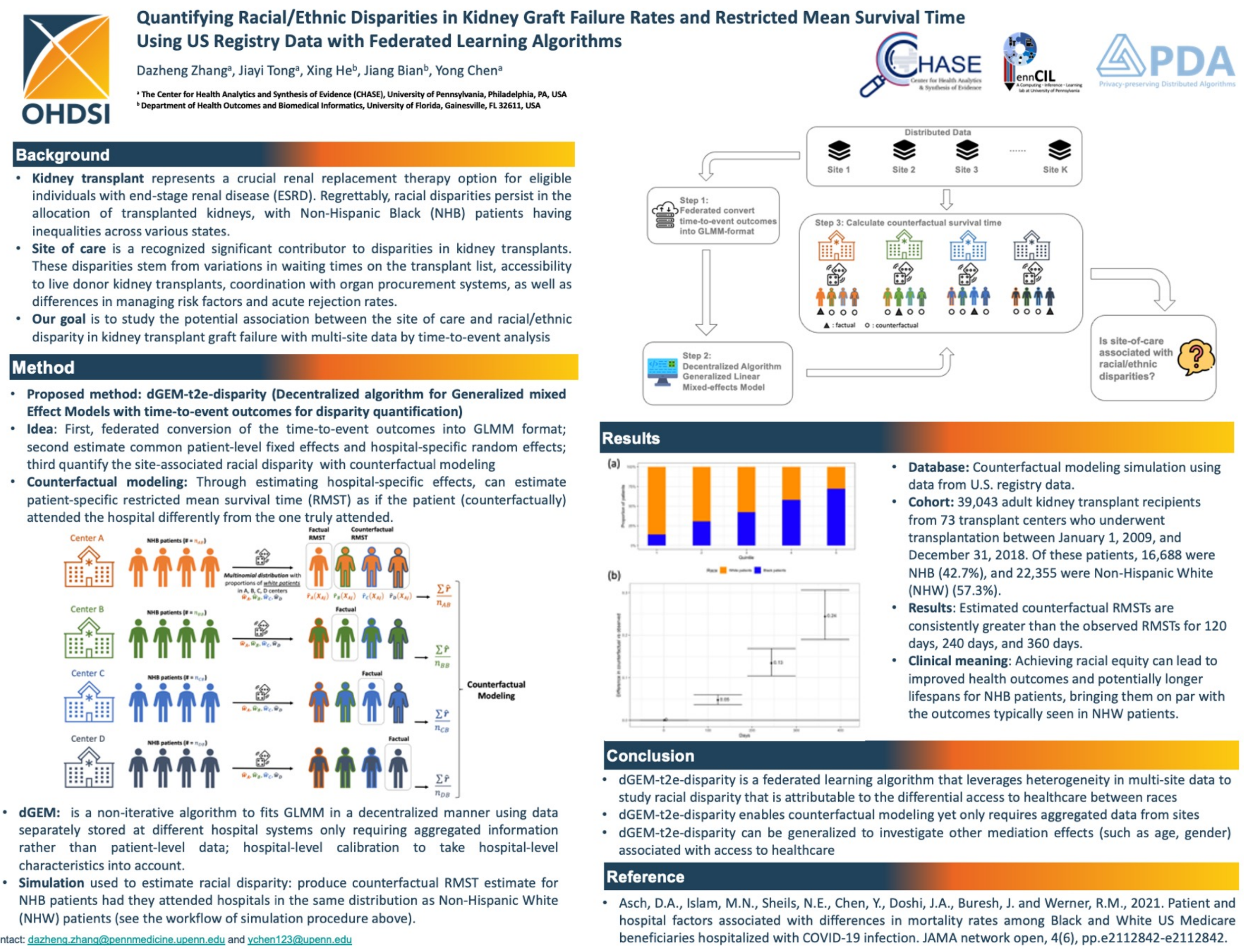
Acknowledgment: This research was supported by the National Institute of Mental Health of the National Institutes of Health under award numbers R01MH129764 and R56MH120826.

#OHDSISocialShowcase This Week

THURSDAY

Quantifying Racial Disparities in Kidney Graft Failure Rates Using US Registry Data with Federated Learning Algorithms

(Dazheng Zhang, Jiayi Tong, Xing He, Jiang Bian, Yong Chen)



Quantifying Racial/Ethnic Disparities in Kidney Graft Failure Rates and Restricted Mean Survival Time Using US Registry Data with Federated Learning Algorithms

Dazheng Zhang[a], Jiayi Tong[a], Xing He[b], Jiang Bian[b], Yong Chen[a]

[a] The Center for Health Analytics and Synthesis of Evidence (CHASE), University of Pennsylvania, Philadelphia, PA, USA
[b] Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL 32611, USA

Background

- **Kidney transplant** represents a crucial renal replacement therapy option for eligible individuals with end-stage renal disease (ESRD). Regrettably, racial disparities persist in the allocation of transplanted kidneys, with Non-Hispanic Black (NHB) patients having inequalities across various states.
- **Site of care** is a recognized significant contributor to disparities in kidney transplants. These disparities stem from variations in waiting times on the transplant list, accessibility to live donor kidney transplants, coordination with organ procurement systems, as well as differences in managing risk factors and acute rejection rates.
- **Our goal** is to study the potential association between the site of care and racial/ethnic disparity in kidney transplant graft failure with multi-site data by time-to-event analysis

Distributed Data
Site 1
Site 2
Site 3
......
Site K
Step 1:
Federated convert time-to-event outcomes into GLMM-format
Step 2:
Decentralized Algorithm Generalized Linear Mixed-effects Model
Step 3: Calculate counterfactual survival time
▲ : factual
○ : counterfactual
Is site-of-care associated with racial/ethnic disparities?

Method

- **Proposed method: dGEM-t2e-disparity (Decentralized algorithm for Generalized mixed Effect Models with time-to-event outcomes for disparity quantification)**
- **Idea**: First, federated conversion of the time-to-event outcomes into GLMM format; second estimate common patient-level fixed effects and hospital-specific random effects; third quantify the site-associated racial disparity with counterfactual modeling
- **Counterfactual modeling:** Through estimating hospital-specific effects, can estimate patient-specific restricted mean survival time (RMST) as if the patient (counterfactually) attended the hospital differently from the one truly attended.

Center A
NHB patients (# = n_{AB})
Multinomial distribution with proportions of white patients in A, B, C, D centers
$\hat{\omega}_A, \hat{\omega}_B, \hat{\omega}_C, \hat{\omega}_D$
Factual RMST
Counterfactual RMST
$\hat{r}_A(X_{Aj})$ $\hat{r}_B(X_{Aj})$ $\hat{r}_C(X_{Aj})$ $\hat{r}_D(X_{Aj})$
$\frac{\sum \hat{r}}{n_{AB}}$
Center B
NHB patients (# = n_{BB})
$\hat{\omega}_A, \hat{\omega}_B, \hat{\omega}_C, \hat{\omega}_D$
Factual
$\frac{\sum \hat{r}}{n_{BB}}$
Center C
NHB patients (# = n_{CB})
$\hat{\omega}_A, \hat{\omega}_B, \hat{\omega}_C, \hat{\omega}_D$
Factual
$\frac{\sum \hat{r}}{n_{CB}}$
Center D
NHB patients (# = n_{DB})
$\hat{\omega}_A, \hat{\omega}_B, \hat{\omega}_C, \hat{\omega}_D$
Factual
$\frac{\sum \hat{r}}{n_{DB}}$
Counterfactual Modeling

- **dGEM:** is a non-iterative algorithm to fits GLMM in a decentralized manner using data separately stored at different hospital systems only requiring aggregated information rather than patient-level data; hospital-level calibration to take hospital-level characteristics into account.
- **Simulation** used to estimate racial disparity: produce counterfactual RMST estimate for NHB patients had they attended hospitals in the same distribution as Non-Hispanic White (NHW) patients (see the workflow of simulation procedure above).

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

Results

- **Database:** Counterfactual modeling simulation using data from U.S. registry data.
- **Cohort:** 39,043 adult kidney transplant recipients from 73 transplant centers who underwent transplantation between January 1, 2009, and December 31, 2018. Of these patients, 16,688 were NHB (42.7%), and 22,355 were Non-Hispanic White (NHW) (57.3%).
- **Results**: Estimated counterfactual RMSTs are consistently greater than the observed RMSTs for 120 days, 240 days, and 360 days.
- **Clinical meaning**: Achieving racial equity can lead to improved health outcomes and potentially longer lifespans for NHB patients, bringing them on par with the outcomes typically seen in NHW patients.

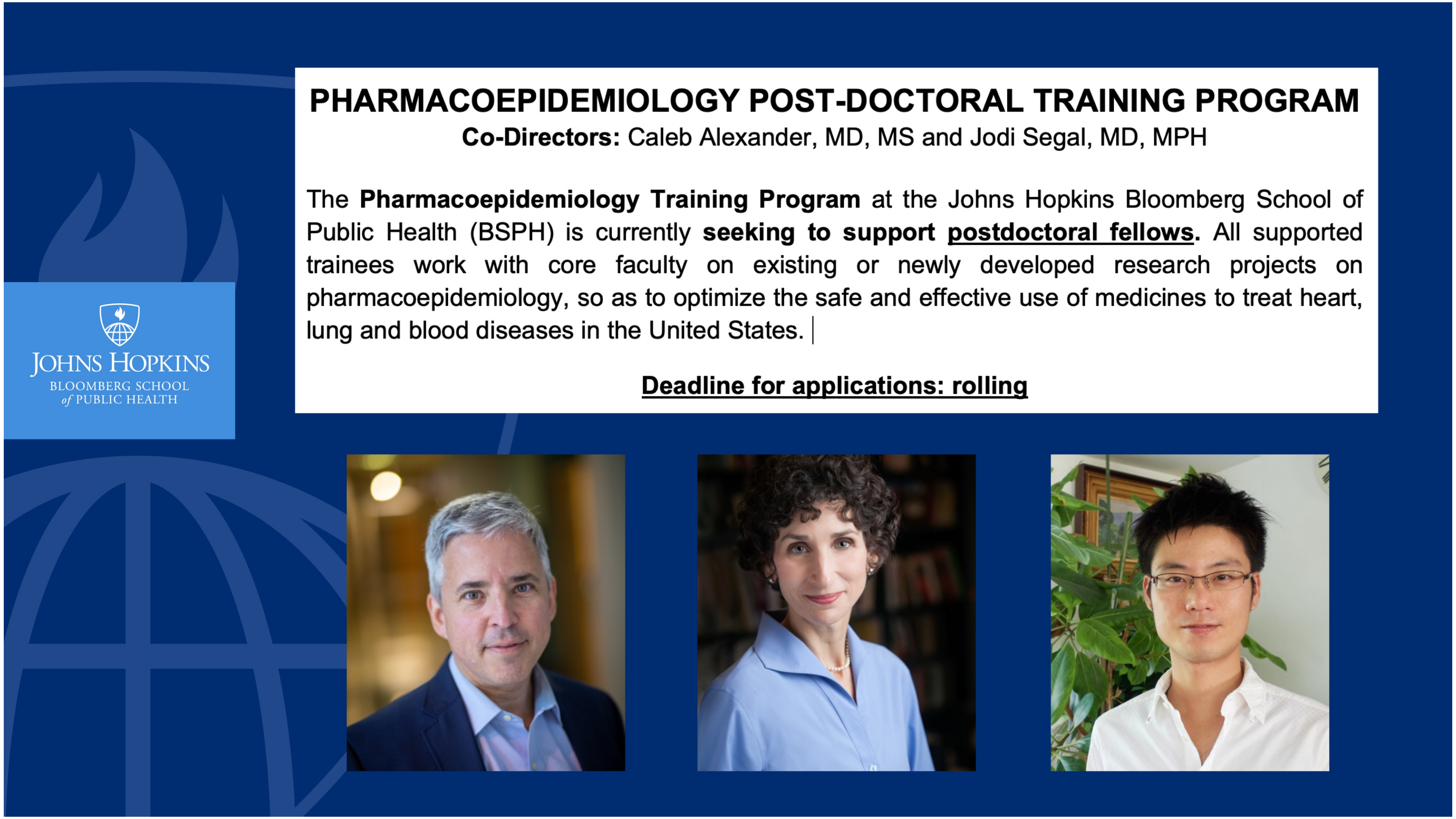
Conclusion

- dGEM-t2e-disparity is a federated learning algorithm that leverages heterogeneity in multi-site data to study racial disparity that is attributable to the differential access to healthcare between races
- dGEM-t2e-disparity enables counterfactual modeling yet only requires aggregated data from sites
- dGEM-t2e-disparity can be generalized to investigate other mediation effects (such as age, gender) associated with access to healthcare

Reference

- Asch, D.A., Islam, M.N., Sheils, N.E., Chen, Y., Doshi, J.A., Buresh, J. and Werner, R.M., 2021. Patient and hospital factors associated with differences in mortality rates among Black and White US Medicare beneficiaries hospitalized with COVID-19 infection. JAMA network open, 4(6), pp.e2112842-e2112842.

Openings: Postdoctoral Fellow, Johns Hopkins Univ.



JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

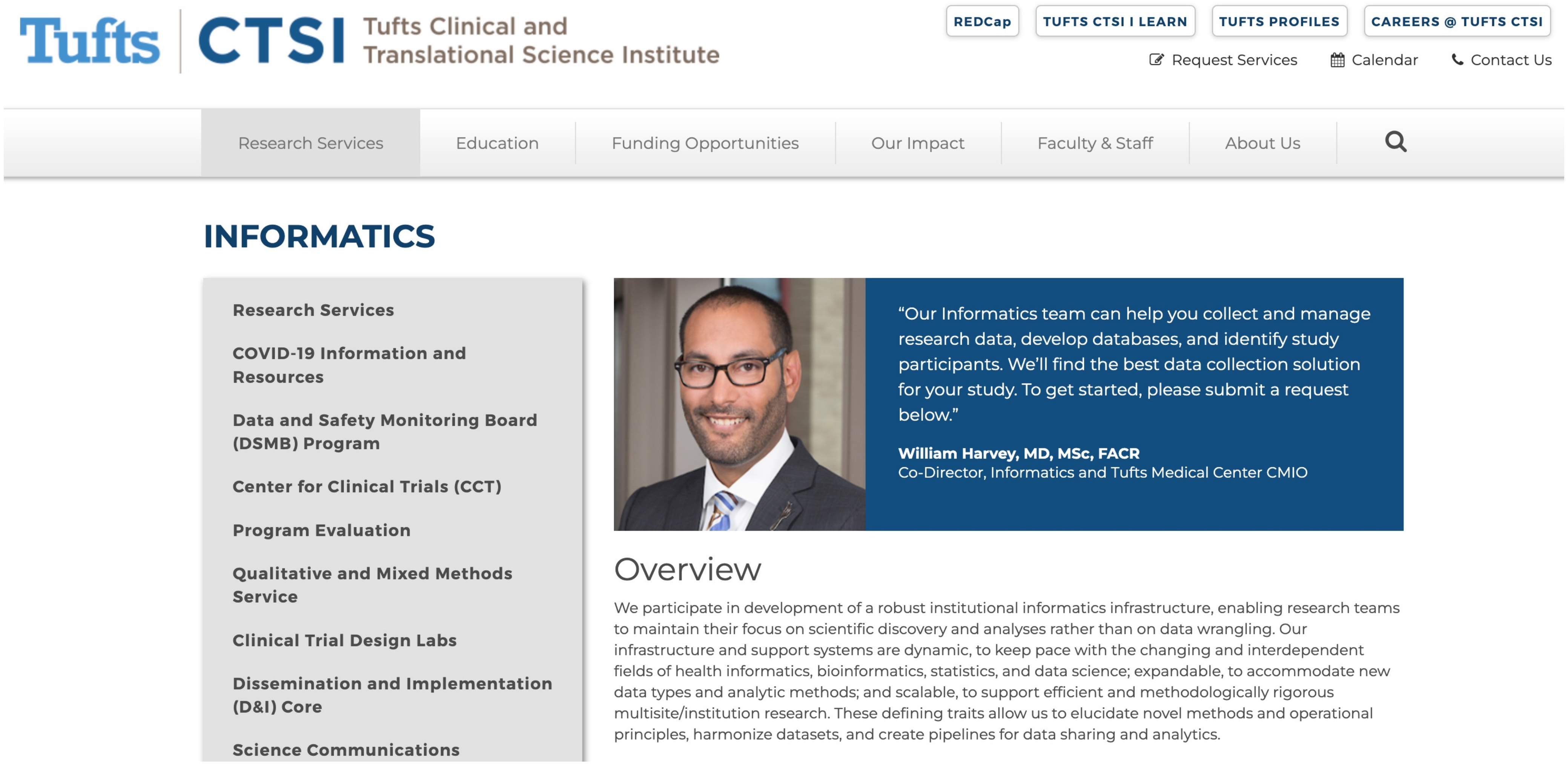
PHARMACOEPIDEMIOLOGY POST-DOCTORAL TRAINING PROGRAM

Co-Directors: Caleb Alexander, MD, MS and Jodi Segal, MD, MPH

The **Pharmacoepidemiology Training Program** at the Johns Hopkins Bloomberg School of Public Health (BSPH) is currently **seeking to support postdoctoral fellows.** All supported trainees work with core faculty on existing or newly developed research projects on pharmacoepidemiology, so as to optimize the safe and effective use of medicines to treat heart, lung and blood diseases in the United States.

Deadline for applications: rolling

Opening: Junior Research Software Engineer, Tufts



Tufts | CTSI Tufts Clinical and Translational Science Institute

REDCap | TUFTS CTSI I LEARN | TUFTS PROFILES | CAREERS @ TUFTS CTSI

Request Services | Calendar | Contact Us

Research Services | Education | Funding Opportunities | Our Impact | Faculty & Staff | About Us

INFORMATICS

- Research Services
- COVID-19 Information and Resources
- Data and Safety Monitoring Board (DSMB) Program
- Center for Clinical Trials (CCT)
- Program Evaluation
- Qualitative and Mixed Methods Service
- Clinical Trial Design Labs
- Dissemination and Implementation (D&I) Core
- Science Communications

> "Our Informatics team can help you collect and manage research data, develop databases, and identify study participants. We'll find the best data collection solution for your study. To get started, please submit a request below."
>
> **William Harvey, MD, MSc, FACR**
> Co-Director, Informatics and Tufts Medical Center CMIO

Overview

We participate in development of a robust institutional informatics infrastructure, enabling research teams to maintain their focus on scientific discovery and analyses rather than on data wrangling. Our infrastructure and support systems are dynamic, to keep pace with the changing and interdependent fields of health informatics, bioinformatics, statistics, and data science; expandable, to accommodate new data types and analytic methods; and scalable, to support efficient and methodologically rigorous multisite/institution research. These defining traits allow us to elucidate novel methods and operational principles, harmonize datasets, and create pipelines for data sharing and analytics.

Where Are We Going?

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

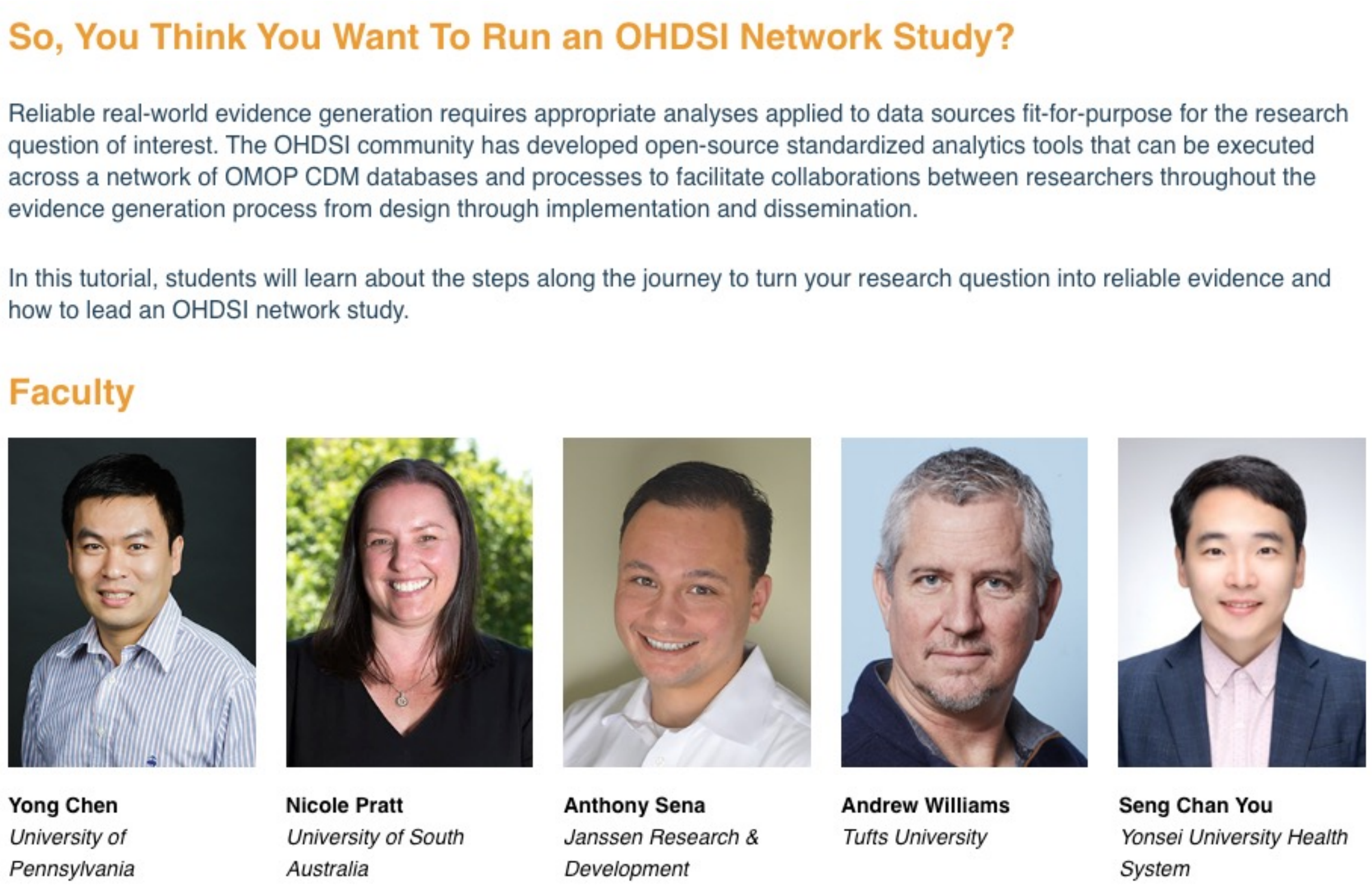
Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

Tutorial: Leading Network Studies



So, You Think You Want To Run an OHDSI Network Study?

Reliable real-world evidence generation requires appropriate analyses applied to data sources fit-for-purpose for the research question of interest. The OHDSI community has developed open-source standardized analytics tools that can be executed across a network of OMOP CDM databases and processes to facilitate collaborations between researchers throughout the evidence generation process from design through implementation and dissemination.

In this tutorial, students will learn about the steps along the journey to turn your research question into reliable evidence and how to lead an OHDSI network study.

Faculty

Yong Chen
University of Pennsylvania

Nicole Pratt
University of South Australia

Anthony Sena
Janssen Research & Development

Andrew Williams
Tufts University

Seng Chan You
Yonsei University Health System

May 21: Open Network Studies

Atif Adam

Associate Director of Epidemiology
IQVIA

Chungsoo Kim

Postdoctoral Associate
Yale University

Linying Zhang

Assistant Professor of Biostatistics
Washington University School of Medicine in St. Louis

Daniel Morales

Wellcome Trust Clinical Research Fellow
University of Dundee

The weekly OHDSI community call is held every Tuesday at 11 am ET.

Everybody is invited!

Links are sent out weekly and available at:
ohdsi.org/community-calls