## Upcoming Community Calls

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 21</td>
<td>Open Studies in the OHDSI Community</td>
</tr>
<tr>
<td>May 28</td>
<td>Collaborator Showcase Brainstorm</td>
</tr>
<tr>
<td>June 4</td>
<td>NO CALL – EUROPEAN SYMPOSIUM</td>
</tr>
<tr>
<td>June 11</td>
<td>European Symposium Review</td>
</tr>
<tr>
<td>June 18</td>
<td>Application of LLMs In Evidence Generation Process</td>
</tr>
<tr>
<td>June 25</td>
<td>Recent OHDSI Publications</td>
</tr>
</tbody>
</table>
Three Stages of The Journey

Where Have We Been?
Where Are We Now?
Where Are We Going?
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.

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 (e.g., 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 robust 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.

NOW 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?
<table>
<thead>
<tr>
<th>Date</th>
<th>Time (ET)</th>
<th>Meeting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wednesday</td>
<td>9 am</td>
<td>OMOP CDM Oncology Outreach/Research Subgroup</td>
</tr>
<tr>
<td>Tuesday</td>
<td>12 pm</td>
<td>Latin America</td>
</tr>
<tr>
<td>Wednesday</td>
<td>3 pm</td>
<td>Joint Vulcan/OHDSI Meeting</td>
</tr>
<tr>
<td>Thursday</td>
<td>9:30 am</td>
<td>Network Data Quality</td>
</tr>
<tr>
<td>Thursday</td>
<td>7 pm</td>
<td>Dentistry</td>
</tr>
<tr>
<td>Friday</td>
<td>9 am</td>
<td>Phenotype Development and Evaluation</td>
</tr>
<tr>
<td>Friday</td>
<td>10 am</td>
<td>GIS-Geographic Information System</td>
</tr>
<tr>
<td>Friday</td>
<td>11:30 am</td>
<td>Clinical Trials</td>
</tr>
<tr>
<td>Friday</td>
<td>11:30 am</td>
<td>Steering Group</td>
</tr>
<tr>
<td>Monday</td>
<td>10 am</td>
<td>Africa Chapter</td>
</tr>
<tr>
<td>Monday</td>
<td>4 pm</td>
<td>Eyecare &amp; Vision Research</td>
</tr>
<tr>
<td>Tuesday</td>
<td>9 am</td>
<td>OMOP CDM Oncology Genomic Subgroup</td>
</tr>
</tbody>
</table>
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

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 Ghebremitew-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.

Career Development
• 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*


May 30, 2024, 11am-12pm EST
Virtually via [Zoom](https://zoom.com)

Please contact Marty Alvarez at [malvarez2@tuftsmmedicalcenter.org](mailto:malvarez2@tuftsmmedicalcenter.org) for calendar invite or questions.
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](https://medicine.utah.edu/dbmi/aime/ai-reliable)

**Organizing Committee**
- Linying Zhang
- Adam Wilcox
- Yves Lussier

**Scientific Program Committee**
- Peter Rijnbeek
- Larry Han
- Xiaoqian Jiang
- Mattia Prosperi
- Xia Ning
- 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
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
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](http://ohdsi.org/OHDSI2024)
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.
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)

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. Assessment of data quality of clinical study data, similar to final data analysis, can also be done across OMOP transformed data. Our project focuses on digital health studies that utilize wearable sensors. Digital health technologies significantly that have been growing recently. 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 all data file presence (all files present for all study events for all data types for all study participants) and all data file content (files adhere to set of rules that investigate data format, data density, feasibility or content consistency).

METHODS

The data quality assessment framework uses Python and is called Sirius. The name was chosen because the Sirius star, being a very bright star, also can appear to be changing colors (this results from reflection, which splits the starlight into the colors of a rainbow). We thought that this was similar to the color coding a result rule (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 18 additional studies. Sirius either evaluates file presence rules or the 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 on 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; that 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 timepoints (OMOP observation table events)

- Consent Info: For studies where consecutive numbering of visits is used (e.g., visit1 instead of absolute dates), 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 files are not converted into format 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 visit each per 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-use 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 XLS) perfect format, for sensors using non-standard output, Sirius function library includes pre-processing action functions that facilitate data conversion and data extraction. For example: it can support reading .bin format of ActiGraph device.

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


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

Jimmy John1, Nick Tatonetti2, Benjamin May2, Nina Bickell3, Parsa Mirhaji3, Surbhi Obeja4, Boudewijn Aasman5, Nina Bickell6, Bruce Rapkin7, Erin M. Henninger8, Pavel Goriacko9, Selvin Soby9
1Montefiore Medicine, 2Columbia University Medical Center, 3Yale 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 in rare disease populations, is a significant challenge. These efforts can lead to health disparities, inequities, and outcomes of care. The DISRUPT project, a collaboration initiative involving Mount Sinai, Columbia University, and the Albert Einstein College of Medicine, aims to address these issues. Supported by the “Break Up to Care” program, DISRUPT seeks to revolutionize the current position of patient trial matching by making cancer clinical trials more 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 NCCT-TCP API and parsing relevant inclusion information through our Parser application; (2) screening existing patient populations for relevant information via a Screen application that leverages our OMOP-CDM database; (3) matching potential patients with eligible patients using our Matching engine.

By leveraging Information technology (IT), the DISRUPT project aims to identify and match understudied 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 industry of potential trial matching, making clinical trials more accessible 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 tool retrieves information from the NCCT-TCP database and parses it into a JSON file. This file contains all the essential information about each clinical trial, including the protocol ID, disease type, stage, and receptor status.

Parser: This tool retrieves information from the NCCT-TCP database via an API and parses it into a JSON file. It extracts essential details such as the protocol ID, disease type, stage, and receptor status for each trial. The parsed information is then formatted for trial matching and stored in an SQLlite DB. However, it’s important to note that the Parser assumes that the target 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, Clarify, etc.) to perform case identification. It takes disease and OMOP-CDM (containing of necessary SQL queries) as inputs and outputs a list of patients classified by cancer type, stage, and receptor status. The Screen application is capable of handling all patients at a low-prognosis of a specific cancer type and anyone with an upcoming appointment in an oncology department in the next few months. The results are divided into two subsets: New Patients and Potential Progressed Patients.

Matcher: This function is an SQL query against the DBs on trial and patients that match. It takes a JSON file as input and outputs a CSV file with a list of potential matches.

This pipeline offers several benefits over traditional methods of identifying and matching patients with clinical trials. It also automates, which saves researchers and clinicians significant amounts 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.

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 identify clinical trial participation by enrolling trials with patients’ needs, rather than relying on fit patients into existing trials.

Contact: Jimmy John, Montefiore-Einstein MD, jjohn@montefiore.org

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

(Praveen Kumar, Christophe G. Lambert)

Abstract

Accurate detection and prevalence estimation of behavioral health conditions, such as opioid use disorder (OUD), is critical for identifying individuals requiring treatment needs. Existing methods struggle with the absence of labeled data, hindering the estimation of the prevalence of OUD and the probability of an individual having OUD. This study introduces a novel framework that leverages both labeled and unlabeled data to estimate the probability of an individual having OUD and the overall population prevalence of OUD using our machine learning algorithm, Positive Unlabeled Learning SelectedTest in Random (PULSMART). The PULSMART algorithm addresses the limitations of traditional methods, which do not accurately reflect the true prevalence of OUD due to the lack of data about OUD cases and the inability to capture the full spectrum of OUD severity.

We applied our machine learning method to healthcare data to analyze the prevalence of OUD and improve our understanding of OUD among individuals in the US. Our results show a significant increase in the prevalence of OUD over time, with the prevalence rate increasing from 0.09% in 2015 to 0.12% in 2020. We also observe a decrease in the proportion of opioid users reporting treatment, from 5.7% in 2015 to 4.5% in 2020.

Discussion and Conclusions

Accurately estimating the prevalence of undiagnosed/unreported behavioral health conditions can have significant implications for public health, including identifying high-risk populations, prioritizing resource allocation, and addressing the negative impacts of these conditions. Our study provides valuable insights into the prevalence and burden of OUD in the US population, highlighting the need for targeted interventions to address this critical public health issue.

Materials and Methods

We applied the PULSMART algorithm to healthcare data to analyze the prevalence of OUD and improve our understanding of OUD among individuals in the US. Our results show a significant increase in the prevalence of OUD over time, with the prevalence rate increasing from 0.09% in 2015 to 0.12% in 2020. We also observe a decrease in the proportion of opioid users reporting treatment, from 5.7% in 2015 to 4.5% in 2020.

Discussion and Conclusions

Accurately estimating the prevalence of undiagnosed/unreported behavioral health conditions can have significant implications for public health, including identifying high-risk populations, prioritizing resource allocation, and addressing the negative impacts of these conditions. Our study provides valuable insights into the prevalence and burden of OUD in the US population, highlighting the need for targeted interventions to address this critical public health issue.

References

(Provide a list of references related to the study, including sources, publication dates, and authors.)

Figure 1: Distribution of opioid use disorder (OUD) prevalence by age group (2015-2020).

Figure 2: Comparison of prevalence of opioid use disorder (OUD) between sexes (2015-2020).
Quantifying Racial/Ethnic Disparities in Kidney Graft Failure Rates and Restricted Mean Survival Time Using US Registry Data with Federated Learning Algorithms

**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)

**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 inequities 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-date by time-to-event analysis

**Method**

- Proposed method: dGEM-12e-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 GLM format; second estimate common patient-level fixed effects and hospital-specific random effects; third quantify the site-associated racial disparity with federated learning.
- 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 they truly attended.

**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-12e-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-12e-disparity enables counterfactual modeling yet only requires aggregated data from sites.
- dGEM-12e-disparity can be generalized to investigate other mediation effects (such as age, gender) associated with access to healthcare.

**Reference**

Openings: Postdoctoral Fellow, Johns Hopkins Univ.

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
INFORMATICS

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

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

Chungsooo Kim
Postdoctoral Associate
Yale University

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: ohdssi.org/community-calls