Collaborator Showcase
Software Demos

OHDSI Community Call
Nov. 21, 2023 • 11 am ET
# Upcoming Community Calls

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov. 21</td>
<td>Showcase Software Demos</td>
</tr>
<tr>
<td>Nov. 28</td>
<td>OHDSI Coordinating Center</td>
</tr>
<tr>
<td>Dec. 5</td>
<td>Recent Publications</td>
</tr>
<tr>
<td>Dec. 12</td>
<td>Happy Birthday OHDSI! Where Have We Come In 10 Years, and in 12 Months?</td>
</tr>
<tr>
<td>Dec. 19</td>
<td>Holiday-Themed Goodbye to 2023!</td>
</tr>
</tbody>
</table>
Three Stages of The Journey

Where Have We Been?
Where Are We Now?
Where Are We Going?
OHDSI Shoutouts!
Dear OHDSI community, I encourage you to participate in the “Saint Nicolas Reindeers” charity this year as well. The idea is that you can send the gifts for kids displaced due to the war in Ukraine. Please see the last year reports in posts above. Although it’s not easy for non-Ukrainian speakers to navigate Ukrainian online stores and organize the delivery, you can send the money to me or Anna,

- My Paypal: dimshitc@gmail.com
- If you don’t have Paypal, you can use Venmo: @Anna-Ostropolets and then we chose kids and send the gifts to the organization.

The beauty of this initiative is that the “Saint Nicolas Reindeers”, the group of Ukrainian volunteers not just deliver presents but organize the whole New Year / Christmas celebration with games and parties for kids.

Of course you can have preferences of what presents and which kids to choose. Then we can work on this together. **Here** you can see the list of letters to the Saint Nicolas with wishes we will fulfill.

Do not hesitate to contact me, if you have any questions: dymshyt@dohdsi.org or in teams.
Three Stages of The Journey

Where Have We Been?
Where Are We Now?
Where Are We Going?
# Upcoming Workgroup Calls

<table>
<thead>
<tr>
<th>Date</th>
<th>Time (ET)</th>
<th>Meeting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wednesday</td>
<td>12 pm</td>
<td>Latin America</td>
</tr>
<tr>
<td>Wednesday</td>
<td>4 pm</td>
<td>Vulcan/OHDSI Meeting</td>
</tr>
<tr>
<td>Friday</td>
<td>9 am</td>
<td>Phenotype Development &amp; Evaluation</td>
</tr>
<tr>
<td>Friday</td>
<td>9 am</td>
<td>GIS – Geographic Information System Development</td>
</tr>
<tr>
<td>Friday</td>
<td>9 am</td>
<td>Vaccine Vocabulary</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>
Global Symposium Homepage

2023 OHDSI Symposium
Oct. 20-22 • East Brunswick, New Jersey

The 2023 OHDSI Global Symposium welcomed more than 400 of our global collaborators together for three days of sharing research, forging new connections and pushing together the OHDSI mission of improving health by empowering a community to collaboratively generate evidence to promote 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.

@OHDSI
www.ohdsi.org
#JoinTheJourney

2023 Global Collaborator Showcase
Observational Data Standards & Management

- FindCQMP: a caositified data network
- Javier Garcia-Tahonera, Porta, Kimokcxova, Pita, Tamaon, Banos, Kulkarni, Dukx, King, Wills, Diet, Kiel, Jointly, Kronen

Global Symposium Homepage

2023 Collaborator Showcase Posters & Software Demos

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

- George Hripcsak, Columbia University
- Ciaran Blackstein, Johnson & Johnson
- Alexander Ouyed, Ophthos Data Services
- Katy Bandrowski, Bootstrap Software
- Peter Rimbock, Ethical MC
- Mengling ‘Mimi’ Peng, National University of Singapore
- Cameron Ross, OHDSI
- Hamed Abdoli, Mpi Research
- Philip Blume, RTI International
- John Matsen, OHDSI

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

Facility will highlight the ways community individuals can participate as well as receive values from the community’s outputs. The course will include topics such as open community data standards — including the CQMP Common Data Model and OHDSI Standardized Vocabularies, open-source analytic tools

Tutorial: Introduction to OHDSI

@OHDSI
www.ohdsi.org
#JoinTheJourney
Conversion of the Canadian Observational Study on Epilepsy (CANOE) REDCap Registry to the OMOP Common Data Model

**INTRO:**
- Epilepsy is a deadly medical condition affecting millions of people worldwide.
- There are many unanswered questions about causes and treatments.
- Many organizations create “home grown” (non-OMOP) patient registries which were not designed with OMOP in mind.
- Widespread use and promotion of the OMOP CDM could result in collaborative research and contribute to our understanding of epilepsy and development of treatments.

**GOALS:**
- Gain better understanding of OMOP CDM
- Use lessons learned to improve future data capture and lead international efforts
- Participate in OHDSI software testing and network studies

**METHODS:**
- Longitudinal REDCap Registry with 6,547 people living with epilepsy
- Clinical-entered data
- Met with international team to discuss common data elements of interest to larger network study
  - Selected minimum viable product (MVP) subset of data
  - Performed ETL (see Figure 1)

**Biggest challenges:**
- Understanding REDCap data structure; evolution
- Identifying validated instruments vs. “homemade” surveys
- Conditions and drug history prior to baseline
  - No start dates, e.g.,
  - Do you take X medication?
  - “1. Yes (current); 2. Yes (past)”
- Training technical team on OMOP tools

**MONDAY**

Conversion of the Canadian Observational Study on Epilepsy (CANOE) REDCap Registry to the OMOP Common Data Model

(Danielle Boyce, Colin Bruce Josephson, Ray Jiang, Samuel Wiebe)

It is possible to ETL a “home grown” patient registry into the OMOP CDM with a multidisciplinary team. Consider the use case and select “minimum viable product” (MVP). Date imputation may be required but must be well documented.
Streamlining cytogenetic data integration for research through ISCN string parsing and storage in OMOP

**TUESDAY**

Streamlining Cytogenetic Data Processing with ISCN Parsing and OMOP

(Ben Smith, Trent Peterson, Jessica Manzyuk)

**INTRODUCTION**
Capturing cytogenetic data for research is often a time-consuming effort built with errors given that source information in most health systems is found on scanned lab reports requiring manual data entry. Human chromosome data is commonly summarized in strings using the standard International System for Human Cytogenomic Nomenclature (ISCN), but it is difficult to capture and process these strings in a format suitable for collaborative research.

Cytogenetic research is needed to improve diagnosis and treatment of numerous cancers. ISCN strings describe genomic rearrangements identified by techniques such as karyotyping, Fluorescence In Situ Hybridization (FISH), and microarrays. Our team explored ways to streamline the process of updating cytogenetic data and integrating it with other clinical information for cancer research. The result involved development of an ISCN string parser and validator that stores the values and metadata in OMOP.

**METHODS**
We studied existing ISCN string parsers, but each was limited in functionality needed for common use cases. We developed a parser and validator using the ANTLR (Another Tool for Language Recognition) programming language and decision tree logic. The parsed results are initially output to JSON for subsequent storage in OMOP.

Within OMOP, we decided to store the captured ISCN strings in the NOTE table and the parsed results in the NOTE_NLP table. Some ISCN strings can be much longer than the 50-character limit found in other OMOP fields that were considered. Additionally, parsed string results align well with the purpose of the NOTE_NLP table.

We developed an accompanying API service that can be integrated with various systems for large-scale processing of ISCN strings. We also developed a simple web client that can be publicly accessed at cyt.principiahealth.

**RESULTS**
Our compilation of source ISCN strings and processed results confirmed that the parser accurately recognized all target abnormalities when complex strings and stored them in appropriate OMOP COM tables—both the string (in NOTE) and the parsed abnormality values (in NOTE_NLP). This data could be analyzed alongside corresponding clinical information for given hypothetic patient records.

**CONCLUSION**
This automated approach to cytogenetic data processing represents a significant opportunity to scale cytogenetic research by reducing manual effort and enable inclusion of cytogenetic data in a wider range of studies. Ideally, researchers will be able to use ATLAS with this parsed data, but currently ATLAS does not work properly with NOTE and NOTE_NLP tables. The parser’s development roadmap includes the addition of optical character recognition (OCR) for extracting ISCN strings directly from scanned lab reports, and extending support beyond traditional karyotyping to include FISH.

Authors: Ben Smith1, Trent Peterson1, Jessica Manzyuk1.
1Principia Health Sciences, Inc.

---

*Principia Health Sciences develops collaborative research environments to engage patients, provide research, and offer collaborations in our quest to discover, develop, and deliver curative therapies*
An initial investigation into more complex stacking methods to improve the transportability of prediction models developed across multiple databases

PRESENTER: Jenna Wong

INTRO AND OBJECTIVES:
• “Stacking” is an ensemble learning method where the predicted probabilities from a set of base learners are used as features to train a meta-learner. Logistic regression is most often used for the meta-learner.
• Stacking was previously used to combine base learners developed in different databases to improve model transportability.
• We investigated if two incremental enhancements could improve the transportability of stacking ensembles: 1) using random forest for the meta-learner, and 2) additionally including age and sex as features in the random forest meta-learner to allow for tailored combinations of base learner predictions by these basic patient features (interactions).

METHODS:
1. We used the OHDSI Patient-Level Prediction framework and sampled up to 500K patients from each of 5 databases: CAEE, MOCO, MOCR, Optum Claims, and Optum Own.
2. We investigated 21 binary outcomes within a target population of people suffering from depression.
3. We iteratively combined two logistic regression base learners from four of the databases and stacked them using x = 20,000; 20,000 observations from the fifth database. The remaining data in the fifth database were used to evaluate performance of the meta-learners and custom models; we evaluated discrimination using the area under the receiver operator characteristic curve (AUC).

RESULTS:
• The logistic regression meta-learners generally performed best (but performed very poorly in a few cases).
• The random forest meta-learners generally performed worse than the logistic regression meta-learners but, unlike the logistic regression meta-learners, were not prone to very poor performance in some cases.
• Adding age and sex to the random forest meta-learner did not have an appreciable impact.
• A custom model often could not be developed when x = 20,000. When x = 20,000, the custom models often performed better than the random forest meta-learners.

CONCLUSIONS:
• Using random forest instead of logistic regression to develop a meta-learner does not improve model transportability.
• When the amount of available training data in a new database is small, developing a stacking ensemble to combine base learners from other databases may be more feasible than developing a new custom model.
• Future work will explore other more parametric approaches for developing complex meta-learners targeting the types of associations and interactions that were of interest in this study.

(Cynthia Yang, Egill Fridgeirsson, Jan Kors, Jenna Reps, Peter Rijnbeek, Ross Williams, Jenna Wong)
Developing a Personalized Clinical Decision Support System for Statin Therapy for Primary Prevention using OMOP-CDM and Deep Learning Techniques

(Su Min Kim, Ju-Hyeon Kim, Yunjin Yum, Eunbeen Jo, Jose Moon, Jong-Ho Kim, Yong Hyun Kim, Eunjeong Ju, Hyung Joon Joo)

Developing a Personalized Clinical Decision Support System for Statin Therapy for Primary Prevention using OMOP-CDM and Deep Learning Techniques

INTRO: Dyslipidemia is considered a major risk factor for cardiovascular diseases (CVD). Statins are the cornerstone of dyslipidemia treatment. However, their clinical use may be limited by side effects like myopathy and an increased risk of diabetes. Current guidelines suggest specific target low-density lipoprotein cholesterol (LDL-C) levels for statin therapy based on patient characteristics but may not account for individual variability in statin response or tolerability. This highlights the need for personalized approaches in statin therapy. This study aimed to develop a deep learning algorithm utilizing OMOP-CDM data to recommend the optimal statin therapy to achieve target LDL-C levels based on individual cardiovascular risk in primary prevention (Figure 1).

METHODS: We extracted data from patients initiating statin therapy for the first time between 2005 and 2017 from the electronic health record databases of three tertiary university hospitals. Exclusions were patients with overt atherosclerotic cardiovascular diseases, those without direct LDL-C measurements, and those unlikely to adhere to statin therapy. A total of 22,447 patient datasets including demographic, clinical information, statin prescription data, diagnostic information, and laboratory data were used. The extracted data were mapped to the OMOP-CDM. This includes mapping local codes to standard OMOP-CDM concepts, and storing them in the appropriate tables (Person, Observation, Drug, Exposure). Data cleaning - Condition ocurrences - Measurement tokens: a total of 16,073 patients from two hospitals and 4,374 patients from another hospital were allocated to the development dataset and the external validation dataset, respectively. Multiple prediction models were developed using the development dataset employing multiple linear regression, KNN, SVM, Random Forests, XGBoost, and Multi-Layer Perceptrons (MLP) with 5-fold cross-validation.

RESULTS: The study findings revealed that a significant proportion of initial statin prescriptions did not achieve the desired low-density lipoprotein cholesterol (LDL-C) treatment level. Highlighting the limitations of empirical prescription practices. Feature importance analysis identified diabetes mellitus, SCCE2/SCCE2-OP, baseline LDL-C level, risk level, and age as the critical features for achieving target LDL-C levels with statin therapy (Figure 2).

CONCLUSION: The results highlight the potential of leveraging standardized OMOP-CDM data and deep learning techniques for personalized statin therapy in dyslipidemia. The scalable CDSS platform developed in this study can be readily implemented, considering the increasing adoption of the OMOP-CDM schema and vocabulary in healthcare institutions. Future work will focus on prospective validation of the algorithm's effectiveness and enhancing scalability by upgrading the platform based on OMOP-CDM.

Take a picture to download the full paper

Su Min Kim, Ju-Hyeon Kim, Yunjin Yum, Eunbeen Jo, Jose Moon, Jong-Ho Kim, Yong Hyun Kim, Eunjeong Ju, Hyung Joon Joo

Contact: Hyung Joon Joo

We are deeply waiting for the collaborators regarding this project as well as the other project using OMOP-CDM.
Prediction of Hospital Length of Stay for Planned Admissions Using OMOP CDM

(Haeun Lee, Seok Kim, Hui-Woun Moon, Se Young Jung, Ho-Young Lee, Sooyoung Yoo)

Background
Accurate prediction of hospital length of stay (LOS) is essential for efficient resource management and healthcare planning. Yet, these models are often compromised by limited covariates and the heterogeneity in healthcare data, hospital systems, and patient populations. This study aims to develop and validate machine learning-based LOS prediction models for planned admissions using the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM).

Methods
Data source:
Electronic health record (EHR) data from Seoul National University Bundang Hospital (SNUBH) in South Korea, converted to the OMOP CDM (version 5.3), were used in the analysis. The database contains the EHR data of 1,903,603 million patients (40,723,280 admissions) accumulated from 2003 to 2020.

Target cohort:
Two cohorts are analyzed: planned admissions and planned admissions with surgical operations. The study included 137,123 planned hospital admissions from January 2016 to December 2020, including 80,180 surgery admissions.

Feature selection:
The covariates included the (1) demographic information, (2) condition occurrence, (3) medication, (4) observation, (5) measurement, (6) procedure, (7) severity index and (8) visit occurrence. Lasso regularization was applied in Logistic regression.

Outcome:
The primary outcome was hospitalization with a length of stay of seven days or more.

Algorithms:
Six algorithms including Logistic regression (LR), random forest (RF), extreme gradient boosting (XGB), light gradient boosting (LGB), gradient boosting (GB) and multilayer perceptron (MLP) algorithms for classification were used to develop the prediction models.

Evaluation:
The performance of the models was evaluated based on the area under the receiver operating characteristic curve (AUC), area under Precision-Recall curve (AUPRC), sensitivity (Recall), specificity, positive predictive value (precision), and accuracy. The Shapley Additive exPlanations (SHAP) analysis was further performed to analyze features importance, while calibration plots were used to assess the reliability of the prediction models.

Conclusions
We demonstrated the use of the OMOP CDM to predict LOS for both planned and surgery admissions. Predictors such as surgical operations, admitting specialty, age have been identified as potential contributors, while time to surgery, severity score, admitting specialty are predictive in surgery admissions. These models could provide insights and strategies for optimizing resource utilization across various healthcare facilities that implement OMOP CDM, potentially leading to reduced surgical mortality rates and improved overall operational efficiency. Additional research necessary to validate the models' performance across different institutions is currently underway, using external validations within the OHDSI network community.
Opening: Limerick Digital Cancer Research Centre

Job Spec

Advertisement/Information for Applicants

Please click on Information for Applicants/Job Description link below for full job

Post Doctoral Researcher (Level 1 or 2) in Cancer Digital Health Real World Evidence (2 Positions)

With over 18,000 students and 2,600 members of staff, the University of Limerick (UL) is an energetic, research led and entering institution with a proud record in innovation and excellence in education, research and scholarship. The dynamic, entrepreneurial and pioneering values which drive UL’s mission and strategy ensure that we capitalise on local, national and international engagement and connectivity. We are renowned for providing an outstanding student experience and conducting leading edge research. Our commitment is to make a difference by shaping the future through educating and empowering our students.

With the River Shannon as a unifying focal point, UL is situated on a superb riverside campus of over 130 hectares. Outstanding recreational, cultural and sporting facilities further enhance the campus’s exceptional learning and research environment.

Applications are invited for the following positions:

Faculty of Education & Health Sciences

School of Medicine

Post Doctoral Researcher (Level 1 or 2) in Cancer Digital Health Real World Evidence (2 Positions) Specific Purpose Contract

Salary Scales: PD1 €42,033 - €48,427 p.a. pro rata

PD2 €49,798 - €54,153 p.a. pro rata

Informal enquiries regarding the post may be directed to:
Professor Aidan Cullen
School of Medicine
University of Limerick
Email: aidan.cullen@ul.ie

"This is a professional training and development role and the training and development relevant to this position will be completed within the period of the contract. Postdoctoral Researchers appointed will be expected to complete the Researcher Career Development Programme."

The closing date for receipt of applications is Friday, 16th December 2022.

Applications must be completed online before 12 noon, Irish Standard Time on the closing date.

The University of Limerick supports blended working
Job Opening: Stanford University

Prospective Postdocs

Open Postdoctoral position, faculty mentor Brian Bateman

Our research team is looking for a postdoctoral scholar in perinatal pharmacoepidemiology. The scholar will work closely with Drs. Brian Bateman and Stephanie Leonard on NIH-funded research projects on the comparative safety and effectiveness of medications in pregnancy and related research topics. Our projects employ advanced analytical methods in large databases, which include claims data and electronic health record data in conventional structures and in common data models. Current topical focus areas include mental health, behavioral health and cardiovascular health of people who are pregnant or postpartum.

Important Info

Faculty Sponsor (Last, First Name):
Bateman, Brian

Other Mentor(s) if Applicable:
Stephanie Leonard

Stanford Departments and Centers:
Anesthes, Periop & Pain Med

Postdoc Appointment Term:
Initial appointment is 1 year with renewal after the first year for an additional 1-2 years by mutual agreement

Appointment Start Date: Flexible start date

Group or Departmental Website:
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?
Nov. 21: Collaborator Showcase Demos

Sarah Gasman
Senior Data Analyst • Boston Medical Center
Leveraging the OMOP Common Data Model to Support Distributed Health Equity Research

Boudewijn Aasman
Senior Data Science Engineer • Montefiore Medical Center
Integrating ATLAS Cohorts with DICOM Images and ECG Waveforms to Enrich Real-World Evidence Research

Laurence Lawrence-Archer
Data Scientist • Odysseus Data Services
Introducing KOIOS: removing impediments in genomic variant identification and mapping

Andrey Soares
Assistant Professor • University of Colorado Anschutz Medical Campus
OMOP-to-BULK FHIR: A tool to convert population level clinical data into standardized FHIR batch data