Collaborations for Strategic Priorities

OHDSI Community Call
Jan. 24, 2023 • 11 am ET

Upcoming OHDSI Community Calls

Date	Topic
Jan. 31	Introduction to Phenotype Phebruary
Feb. 7	Phenotype Phebruary Weekly Update + Workgroup Plans for 2023
Feb. 14	Phenotype Phebruary Weekly Update + Workgroup Plans for 2023
Feb. 21	Phenotype Phebruary Weekly Update + Workgroup Plans for 2023
Feb. 28	Phenotype Phebruary Weekly Update + Workgroup Plans for 2023

Upcoming OHDSI Community Calls

Date	Topic
Jan. 31	Introduction to Phenotype Phebruary
Feb. 7	Phenotype Phebruary Weekly Update + Workgroup Plans for 2023
Feb. 14	Phenotype Phebruary Weekly Update + Workgroup Plans for 2023
Feb. 21	Phenotype Phebruary Weekly Update + Workgroup Plans for 2023
Feb. 28	Phenotype Phebruary Weekly Update + Workgroup Plans for 2023

Jan. 31: Introduction to Phenotype Phebruary

Patrick Ryan

Vice President, Observational Health Data Analytics, Janssen Research and Development, Inc.; Adjunct Assistant Professor, Columbia University

Gowtham Rao

Senior Director, Observational Health Data Analytics, Janssen Research and Development, Inc.; Phenotype Development & Evaluation Workgroup Lead

Azza Shoaibi

Associate Director, Observational Health Data Analytics, Janssen Research and Development, Inc.; OHDSI2022 presenter on “OHDSI Phenotype Phebruary: lessons learned”

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI HADES releases: Characterization v0.0.5

Characterization 0.0.4 | Get started | Reference | Articles | Changelog | HADES

Characterization

Introduction

Characterization is an R package for performing characterization of a target and a comparator cohort.

Features

- Compute time to event
- Compute dechallenge and rechallenge
- Computer characterization of target cohort with and without occurring in an outcome cohort during some time at risk
- Run multiple characterization analyses efficiently
- upload results to database
- export results as csv files

Links

Browse source code

Report a bug

Ask a question

License

Apache License 2.0

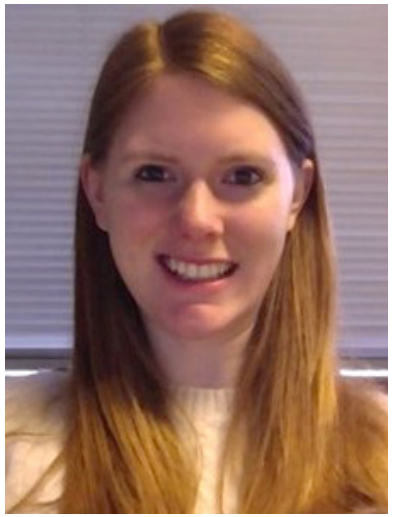
Citation

Citing Characterization

Developers

Jenna Reps
Author, maintainer

Patrick Ryan
Author



New Demos: PatientLevelPrediction v6/Strategus

Jenna Reps, co-lead of the PLP workgroup, recently shared several video tutorials of version 6 of the PatientLevelPrediction tool. The demos are available on both our website and our YouTube page.

Videos

- how to extract data and develop single model using PLP v6
- how to design prediction models and develop multiple models using PLP v6
- demonstrating the PLP v6 shiny app that enables users to interactively explore prediction model results
- how to use the new OHDSI R package Strategus and OHDSI modules to develop an OHDSI prediction development network study
- how to run an OHDSI prediction network study using the new Strategus approach

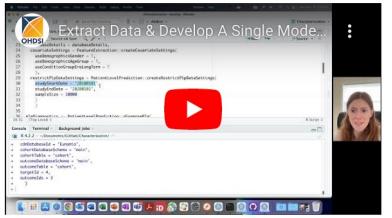
ohdsi.org/plp-v6-demos/

Learn More About Version 6 Of The PatientLevelPrediction Package

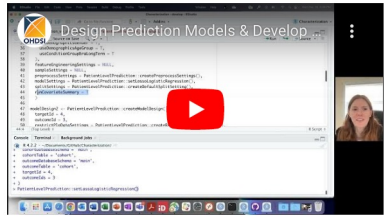
PatientLevelPrediction, a part of the HADES open-source tool library, is an R package for building and validating patient-level predictive models using data in the OMOP Common Data Model format. Check out the PatientLevelPrediction (PLP) github page for more information.

PLP workgroup co-lead and package maintainer Jenna Reps created a series of demo videos to provide assistance with using v6 of the package. You can check out the descriptions and videos here, or on our OHDSI YouTube page (check out the tutorials playlist).

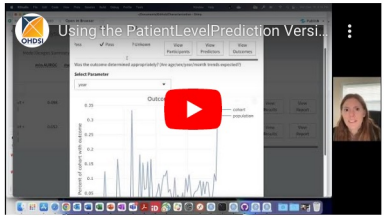
This video demonstrates how to extract data and develop single model using PatientLevelPrediction version 6.



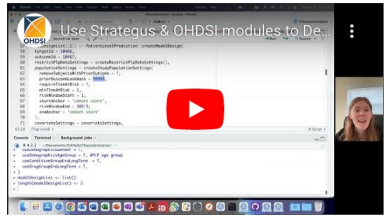
This video demonstrates how to design prediction models and develop multiple models using PatientLevelPrediction version 6.



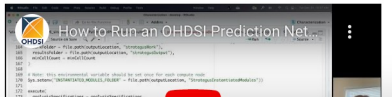
This video demonstrates the PatientLevelPrediction version 6 shiny app that enables users to interactively explore prediction model results.



This video explains how to use the new OHDSI R package Strategus and OHDSI modules to develop an OHDSI prediction development network study. Text instructions are available here.



This video explains explains how to run an OHDSI prediction network study using the new Strategus approach. Text instructions are available here.



OHDSI Shoutouts!

Any shoutouts from the community? Please share and help promote and celebrate OHDSI work!

Have a study published? Please send to sachson@ohdsi.org so we can share during this call and on our social channels.
Let's work together to promote the collaborative work happening in OHDSI!

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Wednesday	7 am	Medical Imaging
Wednesday	11 am	Latin America
Thursday	9:30 am	Data Quality Dashboard
Thursday	7 pm	Dentistry
Friday	9 am	GIS – Geographic Information System General
Friday	9 am	Phenotype Development and Evaluation
Friday	10 am	Education
Friday	11 am	Clinical Trials
Monday	10 am	Healthcare Special Interest Group
Tuesday	9 am	OMOP CDM Oncology Genomic Subgroup

ohdsi.org/workgroups

Save the Date!
October 20-22, OHDSI Global Symposium

Location and more details coming soon

Next CBER Best Seminar

The CBER BEST Seminar Series returns Wed., Feb. 8, at 11 am ET, as 2022 Titan Award recipient **Fan Bu** will provide a presentation on Bayesian Safety Surveillance with Adaptive Bias Correction.

Speaker: Dr. Fan Bu (UCLA)

Description: In this presentation, we will discuss a collaborative project with the FDA CBER BEST Initiative to improve on post-market vaccine safety surveillance procedures through Bayesian sequential analysis. Post-market surveillance on approved vaccine products is essential for addressing safety concerns. The goal is to detect rare or high-risk adverse events that often go undetected in clinical trials due to limited sample sizes. Collaborating with FDA CBER, we have developed a Bayesian alternative surveillance procedure that tackles these challenges in sequential analysis of observational data. The standard statistical approach for surveillance is Maximum Sequential Probability Ratio Test (MaxSPRT). Through comprehensive empirical evaluations on large-scale observational healthcare databases, we show that, compared to MaxSPRT, our Bayesian method offers more flexibility on the surveillance schedule, more transparency and interpretability in decision-making, and better error control through statistical correction of bias in observational data.

Collaboration Opportunity Spotlight: Joint Statistical Meeting (JSM) – Feb. 1 Deadline!

One Community: Informing Decisions and Driving Discovery
TORONTO
Ontario, Canada • August 5–10
JSM2023

Conference Information | Professional Development
Program | Be on the Program
EXPO | Sponsors
Career Service | Home

Follow:

Search

Overview | Participant Guidelines | Sessions | Submissions | Professional Development

Submissions

Invited Session Proposals – Deadline September 8, 2022

Topic-Contributed Session Proposals – Deadline December 8, 2022

Abstract Submissions – Deadline February 1, 2023 NOW OPEN

Presentations may be given on any topic of statistical interest; however, authors are encouraged to submit papers on the theme set by 2023 ASA President Dionne Price, *"One Community: Informing Decisions and Driving Discovery."* Additionally, abstracts with a primary focus on statistical applications are encouraged.

Opens: December 1, 2022
Closes: February 1, 2023

Key Dates

November 15, 2022 - December 8, 2022
Online submission of topic-contributed session proposals

December 1, 2022 - February 1, 2023
Online submission of abstracts (all except invited papers and panels)

January 15, 2023
Computer Technology Workshop (CTW) proposal deadline

January 25, 2023 - April 5, 2023
Online submission of JSM Meeting & Event Requests

January 25, 2023
Deadline to request registration extension for government agencies

May 1, 2023
Registration and housing open

ICPE 2023 Abstract Deadline: Feb. 13

pharmacoepi.org/meetings/annual-conference/

Grant Opportunity

VIEW GRANT OPPORTUNITY

PAR-23-034
NLM Research Grants in Biomedical Informatics and Data Science (R01 Clinical Trial Optional)
Department of Health and Human Services
National Institutes of Health

« Back | Link

Apply | Subscribe

SYNOPSIS | VERSION HISTORY | RELATED DOCUMENTS | PACKAGE

Print Synopsis Details

General Information

Document Type:	Grants Notice	**Version:**	Synopsis 1
Funding Opportunity Number:	PAR-23-034	**Posted Date:**	Oct 06, 2022
Funding Opportunity Title:	NLM Research Grants in Biomedical Informatics and Data Science (R01 Clinical Trial Optional)	**Last Updated Date:**	Oct 06, 2022
		Original Closing Date for Applications:	Jan 07, 2026
Opportunity Category:	Discretionary	**Current Closing Date for Applications:**	Jan 07, 2026
Opportunity Category Explanation:		**Archive Date:**	Feb 12, 2026
Funding Instrument Type:	Grant	**Estimated Total Program Funding:**	
Category of Funding Activity:	Education Health	**Award Ceiling:**	$250,000
		Award Floor:	
Category Explanation:			
Expected Number of Awards:			
CFDA Number(s):	93.310 -- Trans-NIH Research Support 93.879 -- Medical Library Assistance		
Cost Sharing or Matching Requirement:	No		

Upcoming OHDSI APAC Community Calls

Date	Topic
Feb. 16	Training Session #1
Mar. 16	Training Session #2
Apr. 20	Training Session #3
May 18	Training Session #4
June 15	Regional Chapter Mid-Year Updates

Oxford Real World Evidence Summer School

Oxford Summer School 2023: Real World Evidence using the OMOP Common Data Model

COURSE DIRECTORS

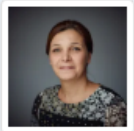
Daniel Prieto-Alhambra

Professor of Pharmaco- and Device Epidemiology

COURSE ADMINISTRATOR

Mahkameh Mafi

Personal Assistant to Professor Prieto-Alhambra



OTHER COURSES

Statistics: Designing clinical research and biostatistics

Brief Description:

Our Real World Evidence Summer School will provide participants with the tools and concepts necessary to plan and execute Real World Evidence studies, with a focus on the use of the OMOP common data model. The course will have morning lectures followed by afternoon practicals where concepts discussed in the morning will be put in practice with hands-on sessions. Practical sessions will have two tracks: a) for those interested in the design of studies and use of existing analytical and data curation tools; and b) for more advanced data scientists and programmers interested in the development or modification of analytical code using R.

Registration: It is now open

Venue: Lady Margaret Hall Talbot Hall Theatre, Norham Gardens, Oxford OX2 6QA

Date: 19th- 23rd June 2023

For booking please use **Booking information**

Please see the Preliminary Programme here

AUDIENCE:

Pharmacists, clinicians, academics (including statisticians, epidemiologists, and related MSc/PhD students); Industry (pharmacy or device) or Regulatory staff with an interest in the use of routinely collected data for research.

LEARNING GOALS:

#OHDSISocialShowcase This Week

Development of Machine Learning models for Cancer Survival among Lung cancer patients with Tyrosine Kinase Inhibitors (TKIs) treatment

Alex PA. Nguyen[1], **Phuc T. Phan**[2], **Min-Huei Hsu**[1], **Jason C. Hsu**[1,2*].
[1] Clinical Data Center, Office of Data Science, Taipei Medical University, Taipei, Taiwan;
[2] International PhD Program in Biotech and Healthcare Management, College of Management, Taipei Medical University, Taipei, Taiwan;



Background

Lung cancer is the most common cause of cancer death worldwide, including in Taiwan. Mutation in the EGFR gene is a driver in lung adenocarcinoma, as this gene is overexpressed in more than 50% of non-small cell lung cancer (NSCLC) in Asia. Most patients benefited from TKI therapies, but 5%-10% of patients did not achieve disease control when administered EGFR-TKIs and therefore acquired drug resistance within 10-12 months.

In this study, we aimed to develop prediction models for lung cancer survival among patients with TKI treatment using a larger number of samples, different data types, and various machine learning algorithms.

Methods

Study Design and Data Source

We conducted a retrospective study in which we obtained the data from the Taiwan Cancer Registry (TCR) database and the Taipei Medical University Clinical Research Database (TMUCRD).

Cohort Selection

Patients with lung cancer (ICD-O-3 code: C33, C34.1) from 2008 to 2018 in the TCR database. Exclusion criteria included individuals under 20, small cell lung cancer (SCLC) patients, and patients who did not receive lung cancer treatment in the three hospitals. Following that, only cancer patients who were undergoing TKIs (i.e., patients using EGFR-TKIs, ATC codes L01EB) were included in our study cohorts.

Outcome Measurement

The outcome of this study was death within two years following diagnosis. Data were censored at the date of death or loss to follow-up, insurance termination, or the study's end on December 31, 2020.

Feature Selection

The selected features were as follows: (1) Demographic information; (2) Cancer condition; (3) Comorbidities; (4) Current medications use; and (5) Laboratory test results. All the features were defined before the time patients were prescribed TKI drugs.

Developing the Machine Learning models

Six machine learning algorithms were used including Logistic Regression (LR), bootstrap aggregation (bagging), gradient boosting machine (GBM), AdaBoost, random forest (RF), and extreme gradient boosting (XGBoost)., to develop the prediction models.

The training set, containing the data of Taipei Medical University Hospital and Wang Fang Hospital. The testing set, including the data of Shuang Ho Hospital, was used to validate the models. The 5-fold cross-validate was applied.

The area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, and F1-score were computed to evaluate and compare the performance of all prediction models.

Contact: **Jason C. Hsu**, International Ph.D. program in Biotech and Health Management, College of Management, Taipei Medical University, Taipei, Taiwan; 11F., No.172-1, Sec. 2, Keelung Rd., Daan Dist., Taipei City 106, Taiwan (R.O.C.); E-mail: *jasohsu@tmu.edu.tw*

Results

Table 1. Baseline demographic of cohort patients in the study

Feature	Training cohort (n=731)	Testing cohort (n=454)
Demographic		
Gender, No. (%)		
Female	327 (44.7%)	206 (45.4%)
Male	404 (55.3%)	248 (54.6%)
Age, Mean (SD), y	67.8 (13.2)	67.3 (12.7)
BMI		
Mean (SD)	23.4 (3.85)	23.2 (4.00)
Median [Min, Max]	23.1 [13.0, 61.3]	22.9 [13.2, 38.1]
Missing	238 (32.6%)	94 (20.7%)
Smoking, No. (%)		
No	356 (48.7%)	222 (48.9%)
Yes	156 (21.3%)	139 (30.6%)
Unknown	219 (30.0%)	93 (20.5%)
Drinking, No. (%)		
No	425 (58.1%)	316 (69.6%)
Yes	85 (11.6%)	45 (9.9%)

Feature	Training cohort (n=731)	Testing cohort (n=454)
Cancer Condition		
Tumor size, No. (%)		
T=<3cm	212 (29.0%)	103 (22.7%)
3=<T=<7cm	321 (43.9%)	210 (46.3%)
T>7cm	63 (8.6%)	48 (10.6%)
Missing	135 (18.5%)	93 (20.5%)
Cancer stage, No. (%)		
stage = 0	52 (7.1%)	28 (6.2%)
stage = 1	17 (2.3%)	9 (2.0%)
stage = 2	81 (11.1%)	33 (7.3%)
stage = 3	547 (74.8%)	368 (81.1%)
stage = 4	34 (4.7%)	16 (3.5%)
Unknown	52 (7.1%)	28 (6.2%)
Mortality, No. (%)	609 (83.3%)	368 (81.1%)

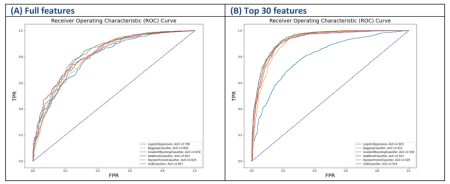


Figure 1. Receiver Operating Characteristic (ROC) Curve of various models

Conclusions

Random forest was observed as the best model when using all features. Moreover, while choosing the top 30 features, Gradient Boosting Classifier was found with the highest AUC of 0.94.

In summary, the model developed using the Gradient Boosting Classifier algorithm had the highest AUC regardless of the mode and was the most suitable tool for NSCLC survival prediction among patients who underwent TKI treatment. In addition, using more types of data (especially laboratory and genomic test results) led to better predictive performance. Cancer stage, cancer size, gender, diagnosis age, and body mass index were the essential features for NSCLC survival prediction.

MONDAY Development of Machine Learning models for Cancer Survival among Lung cancer patients with Tyrosine Kinase Inhibitors (TKIs) treatment (Alex PA. Nguyen, Phuc T. Nguyen, Min-Huei Hsu, Jason C. Hsu)

#OHDSISocialShowcase This Week

Extending the OMOP Standard Vocabulary to Include Botanical Natural Products

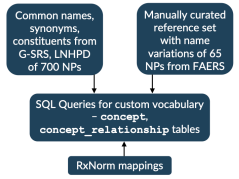
Sanya B. Taneja, Mary F. Paine, Sandra L. Kane-Gill, Richard D. Boyce

INTRODUCTION

OBJECTIVE: extend the OMOP vocabulary to include natural products, their synonyms, phytoconstituents, and name variations to standardize the natural product reports in spontaneous reporting systems.

- Increase in consumption of natural products and/or dietary supplements has led to adverse event concerns.
- Spontaneous reporting systems (e.g., FAERS) can be used for natural product pharmacovigilance by identifying reports with natural products.
- Lack of interoperability in natural product data sources, coverage of synonyms, scientific names and common names, and ambiguity in natural product names are major challenges.

METHODS



RESULTS

- 303 unique natural product Latin binomials
- 2,289 unique concepts in `concept` table
- 2,772 manually curated name variations for 65 natural products from FAERS
- Relationships: *napdi_pt, napdi_is_pt_of, napdi_has_const, napdi_is_const_of, napdi_spell_vr, napdi_is_spell_vr_of, napdi_np_maps_to, napdi_const_maps_to*
- 47,601 reports matched to natural product names, 60,223 reports matched to natural product names & name variations, & 100,522 reports matched to natural product constituents.

303 botanical natural products, 2,289 concepts, and **2,772 name variations** added to **extended OHDSI vocabulary**.

160,745 adverse event reports identified using terms for 65 natural products from the extended vocabulary.

Includes relationships to natural product constituents and RxNorm concepts.

Take a picture to **download** the **full paper**

EXTENDED VOCABULARY

Table 1: `concept` table with green tea concepts.

concept_id	concept_name	vocab_id	concept_class_id
-7000189	Black tea [Camellia sinensis]	NAPDI	Green tea
-7000190	Green tea [Camellia sinensis]	NAPDI	Green tea
-7000191	Oolong tea [Camellia sinensis]	NAPDI	Green tea
-7000192	Tea [Camellia sinensis]	NAPDI	Green tea
-7000193	White Tea [Camellia sinensis]	NAPDI	Green tea
-7000293	Camellia sinensis [Camellia sinensis]	NAPDI	Green tea

Table 2: `concept` table with green tea constituents.

concept_name	constituent_name	concept_id
Green tea	EPICATECHIN	-7001895
Green tea	EPICATECHIN GALLATE	-7002175
Green tea	EPIGALLOCATECHIN	-7001785
Green tea	EPIGALLOCATECHIN GALLATE	-7002248
Green tea	GALLOCATECHIN	-7002061
Green tea	GALLOCATECHIN GALLATE	-7001793

Table 3: `concept` table with green tea name variations.

concept_name	name_variation	concept_id
Green tea	GUARANA GREEN TEA	-7004112
Green tea	CAMELLIA SINENSIS/PANAX GINSENG EXTRACT	-7004069
Green tea	APPLE CIDER VINEGAR + GREEN TEA SUPPLEMENT	-7003800
Green tea	UNSPECIFIED GREEN TEA EXTRACT SUPPLEMENT	-7002714
Green tea	TEA, GREEN (TEA, GREEN)	-7002713

Table 4: Green tea concepts mapped to RxNorm terms.

rxnorm_id	napdi_concept_id	rxnorm_concept	rxnorm_class
19121499	-7001008	GREEN TEA PREPARATION 25 MG	Clinical Drug Comp
1304239	-7001008	GREEN TEA LEAF EXTRACT	Ingredient
1304273	-7001008	GREEN TEA LEAF EXTRACT 1000 MG ORAL TABLET	Clinical Drug
1396861	-7001008	GREEN TEA EXTRACT 315 MG ORAL CAPSULE	Clinical Drug

Details: github.com/dbmi-pitt/np-terminology-imports

FAERS: FDA Adverse Event Reporting System

NaPDI Center of Excellence for Natural Product-Drug Interaction Research

University of Pittsburgh

OHDSI

TUESDAY

Analyzing the Effect of Hypertension on Retinal Thickness Using Radiology Common Data Model (R-CDM) (**Chul Hyoung Park,** Rae Woong Park, Sang Jun Park, Da Yun Lee, Seng Chan You, Ki Hwang Lee)

#OHDSISocialShowcase This Week

Comparing the impact of clean windows across cohorts and databases

PRESENTER: **Rupa Makadia**

INTRO:

- Clean periods of observed person-time allow for the removal of prior exposures (conditions, drugs or procedures) within a cohort.
- The selection of this time can vary based on design of the study, prior knowledge, or random assignment by researchers.
- In this study we examine the trade-offs between various time-windows for a clean window to identify new events within a phenotype across a variety of databases.

METHODS:

- *Database*
  - IBM MarketScan® Databases Commercial Claims (CCAE)
- *Phenotypes (10)*
  - Acute myocardial infarction, myocarditis/pericarditis, deep vein thrombosis, pulmonary embolism, disseminated intravascular coagulation, non-hemorrhagic stroke, hemorrhagic stroke, cerebral venous thrombosis, peripheral arterial thrombosis, and thrombocytopenia
- *Clean window times (0, 14, 28, 90, 180, 270, 365)*
  - Clean windows applied during cohort creation in ATLAS, 70 cohorts were created, 10 for each phenotype
  - Custom code was created to summarize the number of persons and events across each time window.

RESULTS:

- Table 1 presents the number of patients identified (next to phenotype name), with the counts of persons by events (max=10). The total persons represents the total patients identified with 10 events.
- Phenotypes restricted by inpatient are banded in black.

Clean windows should be empirically derived in cohort studies to eliminate prevalent cases for identification of new events

Table 1. Phenotypes by clean window and number of persons by event with a maximum of 10.

		Events											
	Days	1	2	3	4	5	6	7	8	9	10	Total events	Total persons
Acute myocardial infarction (n=527,411)	0	371,577	92,449	29,657	13,566	7,670	4,176	2,314	1,296	773	513	871,671	523,991
	14	447,805	53,971	14,315	6,462	3,014	1,046	396	166	86	39	652,772	527,300
	28	462,567	48,387	11,280	3,466	1,075	335	141	52	32	23	617,118	527,358
	90	488,661	32,747	4,648	927	278	93	33	12	4	3	574,207	527,406
	180	496,228	27,156	3,275	567	141	32	7	2	2	1	563,623	527,411
	270	500,466	23,946	2,550	349	82	14	3	1	-	-	557,928	527,411
	365	503,873	21,280	1,984	222	45	4	3	-	-	-	553,543	527,411
Cerebral venous thrombosis (n=3,329)	0	2,212	732	215	79	29	22	17	8	3	2	5,334	3,319
	14	3,002	246	54	18	5	1	1	1	-	1	3,784	3,329
	28	3,081	199	43	3	3	-	-	-	-	-	3,635	3,329
	90	3,201	113	13	1	1	-	-	-	-	-	3,475	3,329
	180	3,259	63	7	-	-	-	-	-	-	-	3,406	3,329
	270	3,281	46	2	-	-	-	-	-	-	-	3,379	3,329
	365	3,299	30	-	-	-	-	-	-	-	-	3,359	3,329
Disseminated intravascular coagulation (n=22,667)	0	14,900	4,968	1,428	585	307	159	86	65	34	33	37,424	22,565
	14	20,751	1,577	243	56	18	9	7	-	2	2	25,140	22,665
	28	21,667	864	101	22	4	4	2	1	1	-	23,877	22,666
	90	22,284	346	32	3	1	1	-	-	-	-	23,095	22,667
	180	22,454	194	17	1	-	1	-	-	-	-	22,903	22,667
	270	22,517	139	10	1	-	-	-	-	-	-	22,829	22,667
	365	22,557	104	6	-	-	-	-	-	-	-	22,783	22,667
Non-hemorrhagic stroke (n=379, 953)	0	251,108	64,854	24,108	12,649	7,734	4,998	3,501	2,453	1,865	1,330	738,190	374,600
	14	314,766	42,913	12,334	4,665	2,266	1,174	657	367	238	165	491,947	379,545
	28	326,811	37,949	9,331	3,165	1,298	646	301	178	92	57	460,357	379,828
	90	344,836	28,114	5,076	1,283	399	156	54	23	6	2	425,035	379,949
	180	353,370	22,441	3,248	663	160	49	18	3	-	1	411,902	379,953
	270	358,314	18,905	2,251	377	79	22	2	2	1	-	404,951	379,953
	365	361,954	16,072	1,657	223	41	2	2	2	-	-	400,208	379,953
Hemorrhagic stroke (n=123,178)	0	73,152	25,413	9,750	5,067	2,978	1,881	1,215	887	598	454	255,200	121,395
	14	106,133	12,451	2,799	914	397	181	102	64	40	23	149,132	123,104
	28	111,691	8,977	1,677	456	181	81	42	19	21	11	138,957	123,156
	90	117,829	4,747	497	73	20	6	3	3	-	-	129,287	123,178
	180	119,979	2,975	187	33	3	1	-	-	-	-	126,643	123,178
	270	120,886	2,171	106	13	2	-	-	-	-	-	125,608	123,178
	365	121,639	1,660	71	8	-	-	-	-	-	-	125,004	123,178
Thrombocytopenia (n=1,157,870)	0	578,347	208,738	99,595	59,930	40,300	29,104	21,767	16,741	13,279	10,814	4,406,172	1,078,615
	14	706,446	184,179	86,318	50,807	32,549	21,959	15,776	11,753	8,687	6,727	2,768,501	1,125,201
	28	758,373	179,997	80,245	44,475	26,978	17,424	12,181	8,675	6,223	4,598	2,332,170	1,139,169
	90	861,842	161,852	61,145	29,488	16,197	9,512	5,854	3,765	2,511	1,719	1,789,701	1,153,885
	180	939,171	141,316	43,945	17,602	7,896	3,747	1,881	1,053	572	277	1,520,669	1,157,460
	270	990,580	122,008	30,294	9,422	3,371	1,282	508	242	76	51	1,394,838	1,157,834
	365	1,023,097	105,695	21,311	5,522	1,576	448	137	63	15	4	1,332,726	1,157,868
Deep vein thrombosis (n=785, 378)	0	346,081	119,529	67,203	46,311	33,643	25,566	20,049	16,365	13,588	11,351	3,880,684	699,686
	14	432,790	123,825	68,464	42,994	28,039	19,457	14,002	10,008	7,697	5,978	2,402,410	753,254
	28	478,515	130,111	64,803	35,560	21,333	13,923	9,597	6,724	5,100	3,925	1,838,306	769,591
	90	586,705	117,555	41,452	18,624	9,949	5,414	3,236	1,944	1,266	824	1,180,260	786,969
	180	652,604	91,492	25,796	9,695	4,044	1,855	847	409	192	91	995,882	787,025
	270	686,435	74,843	17,273	5,007	1,673	585	191	60	25	4	921,958	786,096
	365	705,876	63,716	11,998	2,834	729	169	46	8	2	-	885,701	785,378
Myocarditis/pericarditis (n=139,677)	0	87,723	25,211	10,173	5,496	3,296	2,019	1,402	964	705	510	297,728	137,499
	14	106,185	18,169	6,762	3,174	1,779	1,068	718	483	359	261	222,873	138,958
	28	112,690	15,591	5,340	2,368	1,317	776	496	336	271	168	200,097	139,353
	90	122,477	11,337	3,175	1,314	638	376	245	135	69	40	170,080	139,806
	180	127,800	8,707	2,017	728	295	136	51	32	15	12	157,411	139,793
	270	131,075	6,862	1,272	362	94	30	16	4	-	2	151,008	139,737
	365	133,007	5,568	841	187	65	13	4	2	-	-	147,811	139,677
Pulmonary embolism (n=437,437)	0	162,867	59,389	36,220	27,398	21,669	17,222	13,921	11,668	9,584	8,426	2,801,205	368,364
	14	216,679	63,785	40,001	28,062	20,019	14,390	10,391	7,814	5,892	4,740	1,626,995	411,773
	28	245,244	70,931	40,841	26,542	15,286	10,075	6,934	5,042	3,865	2,810	1,175,729	425,570
	90	317,798	68,921	25,785	11,826	6,273	3,550	1,973	1,247	791	481	680,223	438,645
	180	361,673	51,753	15,251	5,600	2,287	1,101	464	213	101	50	558,208	438,493
	270	383,543	41,139	9,323	2,637	834	282	102	28	8	4	511,250	437,900
	365	395,688	33,736	6,210	1,391	306	82	21	2	1	-	489,548	437,437
Peripheral arterial thrombosis (n=69,936)	0	-	44,878	-	12,333	-	4,809	-	2,465	-	1,538	304,188	66,023
	14	51,397	10,547	3,732	1,761	970	520	322	198	159	111	111,992	69,717
	28	54,017	9,575	3,154	1,423	740	406	236	152	122	73	102,496	69,898
	90	58,643	7,745	2,149	829	377	161	92	35	20	11	88,141	70,062
	180	61,873	6,300	1,319	373	117	32	8	1	-	1	80,784	70,024
	270	63,923	5,030	826	158	34	5	-	-	-	-	77,293	69,976
	365	65,186	4,118	545	74	13	-	-	-	-	-	75,418	69,936

RESULTS:

- 94% (496,228) of people have a single acute myocardial infarction diagnosis within 180 days whereas only 70% (371,577) people have a single acute myocardial infarction diagnosis with no clean window.
- Phenotypes that occur in any place show the most variability of events compared to those phenotypes with an inpatient restriction.
- Rarer events (cerebral venous thrombosis, disseminated intravascular coagulation) encapsulate most events within 90 days.

CONCLUSIONS:

- **This analysis highlights the need to consider this diagnostic when building a cohort. The results from these phenotypes show wide variability among them and highlight the need for careful consideration in a study design.**
- **There is a possibility that new events maybe missed especially in cases such as deep vein thrombosis where the clean window is 365 days.**
- **Plausibility of events occurring in observational data and identification of new events or continuation of care from a prior event should be carefully evaluated for conditions that could have reoccurrence.**

Rupa Makadia[1], Kevin Haynes[1], Patrick Ryan[1]

[1]Janssen Research & Development, Titusville, NJ

OHDSI

Johnson & Johnson

OFFICE OF THE CHIEF MEDICAL OFFICER

WEDNESDAY Comparing the impact of clean windows across cohorts and databases (Rupa Makadia, Kevin Haynes, Patrick Ryan)

#OHDSISocialShowcase This Week

Assessing and Benchmarking Data Quality and Diversity in the *All of Us* Research Program

Lina Sulieman, PhD;[1] Karthik Natarajan[2], PhD; Kayla Marginean, MS;[1] Robert Carroll, PhD;[1] Paul Harris, PhD[1]
[1]Department of Biomedical Informatics, Vanderbilt University Medical Center; [2]Department of Biomedical Informatics, Columbia University, New York, NY

Objective

Assessing the data quality of Electronic Health Records (EHR) in the *All of Us* Research Program using OHDSI tools and libraries

Background/Introduction

- Quantifying and benchmarking data quality in research clinical repositories is crucial
  - Ensuring the utility of the data
  - Improving the reproducibility and research credibility of research
- Quality assessment methods:
  - Calculating clinical data quality metrics: completeness, plausibility, conformance
  - Replication: replicating existing clinical studies (e.g., phenotype algorithms)
  - Benchmarking against expected published numbers
- The *All of Us* Research Program is a national initiative collecting Electronic Health Records, surveys, and genetic data
- The *All of Us* Research Program collects data from over 50 EHR sites
- Aim: Assessing the *All of Us* EHR data quality by quantifying: completeness, plausibility, conformance, and the prevalence of phenotypes

Methods

- Dataset: *All of Us* controlled tier launched in March 2022, we extracted participants who have EHR
- Quality assessment methods:
  - Phenotype replications:
    - Assessing the prevalence of phenotypes
    - Implemented OHDSI phenotype library to extract 212 phenotypes using 763 algorithms
  - Benchmarking: Comparing the prevalence of phenotypes in *All of Us* to the prevalence reported by the Center for Disease and Control (CDC)
  - Quality metrics: Running OHDSI data quality dashboard package and other ad-hoc metrics to calculate the following:
    - Plausibility: the extent to which the values agree with internal and external knowledge
    - Conformance: the percentage of the dataset that complies with standards and constraints
    - Completeness: the percentage of data that is expected to be present
    - Additional completeness metrics: the percentage of participants who have core measurements: height, weight, Body Mass Index (BMI), cholesterol, and heart rate per EHR site

Results

- Dataset: *All of Us* Research Program controlled tier included 331,382 participants
  - 76% were considered underrepresented in biomedical research
  - White participants: 55.81% compared to 76.30% white in the US general population
  - Female participants: 60%
  - Have any EHR data: 224,507 (67.7%) participants

Results

- Replication (Figure 1):
  - 223,018 participants with at least one of the 212 EHR-based phenotypes
  - Racial distributions varied: 413 (55%) of the cohorts had 60% or lower white participants
- Benchmarking:
  - Mean prevalence of the US leading causes of death phenotypes: close to or slightly higher than the prevalence reported by CDC (except for Alzheimer's disease, suicide, influenza) as Figure 2 shows
- Quality metrics:
  - OHDSI quality dashboard (Figure 3)
  - Common quality problems:
    - Plausibility: out-of-range values, gender-specific conditions and procedures concepts
    - Conformance: Non-standard/non-existing
    - Completeness: Null or zero values
  - Core measurement completeness per site measured by the percentage of participants:
    - Height: 72.0% (3%-100%)
    - Weight: 76.5% (3%-100%)
    - BMI: 76.8% (11%-100)
    - Heart rate: 80.5% (11%-100)

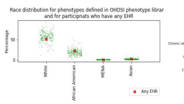


Figure 1. Race distribution in phenotype algorithms applied on the *All of Us* Research Program dataset compared to the race distribution for participants with any EHR

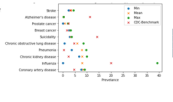


Figure 2. The prevalence of phenotypes that are the leading causes of death as reported by the CDC. We reported the minimum, maximum, and mean prevalence for all algorithms per phenotype

	Verification			Validation			Total		
	Pass	Fail	Total	Pass	Fail	Total	Pass	Fail	Total
Plausibility	2611	202	2813	70%	0	279	107	0%	2509
Conformance	492	189	681	72%	92	92	104	88%	584
Completeness	209	87	296	77%	13	3	25	87%	222
Total	3612	468	3980	89%	113	293	406	28%	3725

Figure 3. The plausibility, conformance, and completeness metrics in the *All of Us* Research Program

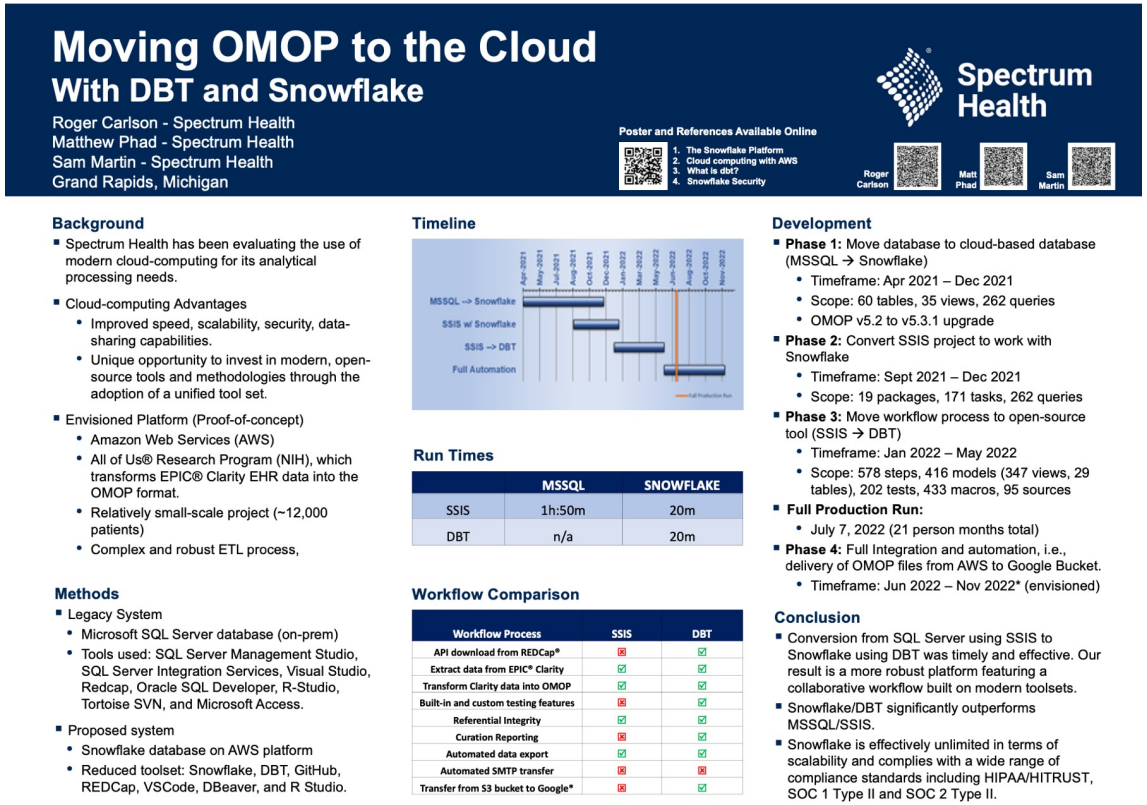
Conclusions

- Data quality and diversity are essential factors that can improve clinical research reproducibility in major clinical research repositories such as *All of Us* Research Program
- We used OHDSI tools to assess data quality using: replication, benchmarking, and quality dimension metrics (plausibility, conformance, completeness)
- Our analysis demonstrated the diversity of the *All of Us* EHR
  - Non-white participants' representation is high compared to other research repositories
  - Quality aspects to further investigate:
  - Identifying possible reasons for having different prevalence values compared to CDC
    - Simpler versions of phenotype algorithms
    - Site recruitment might influence the disease prevalence within *All of Us* Research Program (recruiting from breast cancer clinic)
  - Low conformance and core measurement completeness: mapping differences in EHR sites

Email: lina.Sulieman@vumc.org
: @LinaSulieman

THURSDAY Assessing and Benchmarking Data Quality and Diversity in the All of Us (Lina Sulieman, Karthik Natarajan, Kayla Marginean, Robert Carroll, Paul Harris)

#OHDSISocialShowcase This Week



Moving OMOP to the Cloud
With DBT and Snowflake

Roger Carlson - Spectrum Health
Matthew Phad - Spectrum Health
Sam Martin - Spectrum Health
Grand Rapids, Michigan

Poster and References Available Online

1. The Snowflake Platform
2. Cloud computing with AWS
3. What is dbt?
4. Snowflake Security

Spectrum Health

Roger Carlson | Matt Phad | Sam Martin

Background

- Spectrum Health has been evaluating the use of modern cloud-computing for its analytical processing needs.
- Cloud-computing Advantages
  - Improved speed, scalability, security, data-sharing capabilities.
  - Unique opportunity to invest in modern, open-source tools and methodologies through the adoption of a unified tool set.
- Envisioned Platform (Proof-of-concept)
  - Amazon Web Services (AWS)
  - All of Us® Research Program (NIH), which transforms EPIC® Clarity EHR data into the OMOP format.
  - Relatively small-scale project (~12,000 patients)
  - Complex and robust ETL process,

Methods

- Legacy System
  - Microsoft SQL Server database (on-prem)
  - Tools used: SQL Server Management Studio, SQL Server Integration Services, Visual Studio, Redcap, Oracle SQL Developer, R-Studio, Tortoise SVN, and Microsoft Access.
- Proposed system
  - Snowflake database on AWS platform
  - Reduced toolset: Snowflake, DBT, GitHub, REDCap, VSCode, DBeaver, and R Studio.

Timeline

Run Times

	MSSQL	SNOWFLAKE
SSIS	1h:50m	20m
DBT	n/a	20m

Workflow Comparison

Workflow Process	SSIS	DBT
API download from REDCap®	☒	☑
Extract data from EPIC® Clarity	☑	☑
Transform Clarity data into OMOP	☑	☑
Built-in and custom testing features	☒	☑
Referential Integrity	☑	☑
Curation Reporting	☒	☑
Automated data export	☑	☑
Automated SMTP transfer	☒	☒
Transfer from S3 bucket to Google*	☒	☑

Development

- **Phase 1:** Move database to cloud-based database (MSSQL → Snowflake)
  - Timeframe: Apr 2021 – Dec 2021
  - Scope: 60 tables, 35 views, 262 queries
  - OMOP v5.2 to v5.3.1 upgrade
- **Phase 2:** Convert SSIS project to work with Snowflake
  - Timeframe: Sept 2021 – Dec 2021
  - Scope: 19 packages, 171 tasks, 262 queries
- **Phase 3:** Move workflow process to open-source tool (SSIS → DBT)
  - Timeframe: Jan 2022 – May 2022
  - Scope: 578 steps, 416 models (347 views, 29 tables), 202 tests, 433 macros, 95 sources
- **Full Production Run:**
  - July 7, 2022 (21 person months total)
- **Phase 4:** Full Integration and automation, i.e., delivery of OMOP files from AWS to Google Bucket.
  - Timeframe: Jun 2022 – Nov 2022* (envisioned)

Conclusion

- Conversion from SQL Server using SSIS to Snowflake using DBT was timely and effective. Our result is a more robust platform featuring a collaborative workflow built on modern toolsets.
- Snowflake/DBT significantly outperforms MSSQL/SSIS.
- Snowflake is effectively unlimited in terms of scalability and complies with a wide range of compliance standards including HIPAA/HITRUST, SOC 1 Type II and SOC 2 Type II.

FRIDAY Moving OMOP to the cloud with DBT and Snowflake (Roger Carlson, Matthew Phad, Samuel Martin)

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?

Collaborations for Strategic Opportunities

Anna Ostropolets

Data Scientist, Odysseus Data Services, Inc.

PhD Graduate, Columbia University

Clair Blacketer

Director, Janssen Research and Development, Inc.

Patrick Ryan

Vice President, Observational Health Data Analytics, Janssen Research and Development, Inc.; Adjunct Assistant Professor, Columbia University