What Can OHDSI Do Together in 2025?

OHDSI Community Call
Jan. 7, 2025 • 11 am ET

OHDSI
OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Upcoming Community Calls

Date	Topic
Jan. 7	What Can OHDSI Go In 2025?
Jan. 14	Connnections for Future Collaborations
Jan. 21	Clinical Guideline Review, Session I
Jan. 28	Clinical Guideline Review, Session II
Feb. 4	First Week of 2025 Workgroup OKRs/Phenotype Phebruary

Three Stages of The Journey

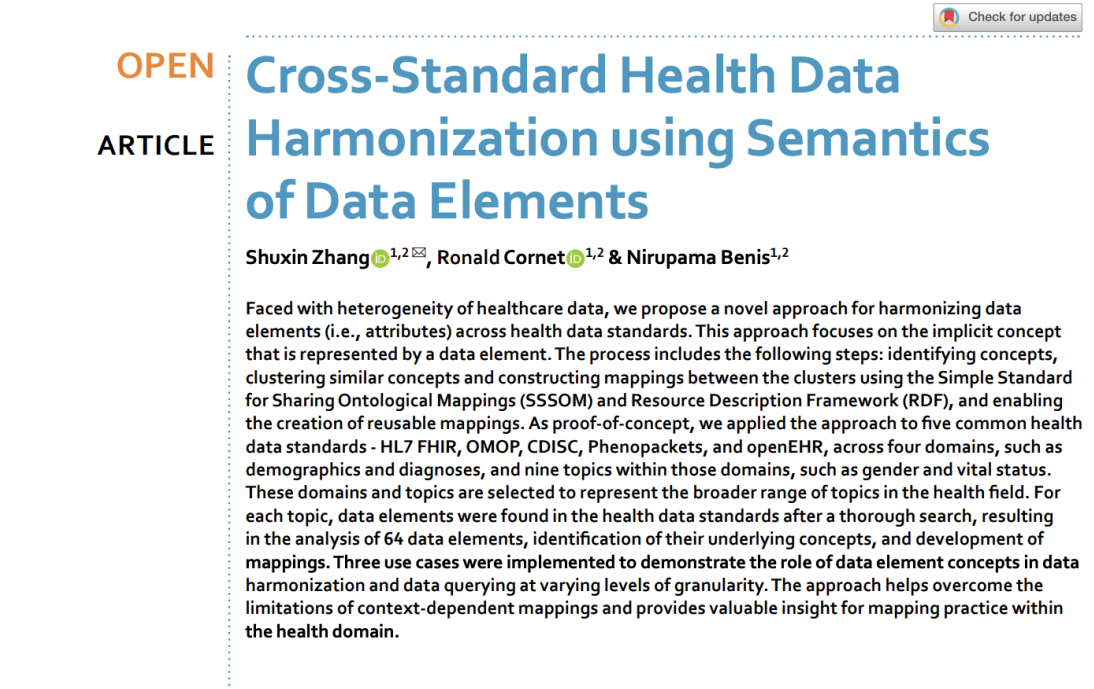
Where Have We Been?

Where Are We Now?

Where Are We Going?

OHDSI Shoutouts!

Congratulations to the team of **Shuxin Zhang, Ronald Cornet and Nirupama Benis** on the publication of **Cross-Standard Health Data Harmonization using Semantics of Data Elements** in *Scientific Data.*



scientific data

Check for updates

OPEN

ARTICLE

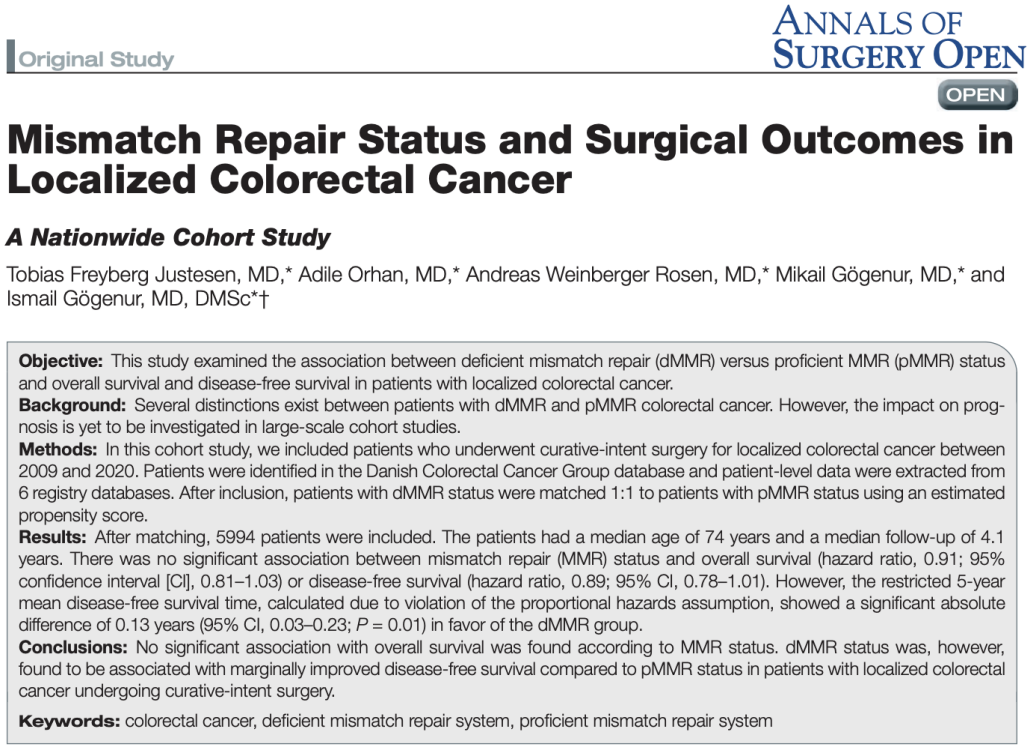
Cross-Standard Health Data Harmonization using Semantics of Data Elements

Shuxin Zhang[1,2]✉, Ronald Cornet[1,2] & Nirupama Benis[1,2]

Faced with heterogeneity of healthcare data, we propose a novel approach for harmonizing data elements (i.e., attributes) across health data standards. This approach focuses on the implicit concept that is represented by a data element. The process includes the following steps: identifying concepts, clustering similar concepts and constructing mappings between the clusters using the Simple Standard for Sharing Ontological Mappings (SSSOM) and Resource Description Framework (RDF), and enabling the creation of reusable mappings. As proof-of-concept, we applied the approach to five common health data standards - HL7 FHIR, OMOP, CDISC, Phenopackets, and openEHR, across four domains, such as demographics and diagnoses, and nine topics within those domains, such as gender and vital status. These domains and topics are selected to represent the broader range of topics in the health field. For each topic, data elements were found in the health data standards after a thorough search, resulting in the analysis of 64 data elements, identification of their underlying concepts, and development of mappings. Three use cases were implemented to demonstrate the role of data element concepts in data harmonization and data querying at varying levels of granularity. The approach helps overcome the limitations of context-dependent mappings and provides valuable insight for mapping practice within the health domain.

OHDSI Shoutouts!

Congratulations to the team of **Tobias Freyberg Justesen, Adile Orhan, Andreas Weinberger Rosen, Mikail Gögenur, and Ismail Gögenur** on the publication of **Mismatch Repair Status and Surgical Outcomes in Localized Colorectal Cancer: A Nationwide Cohort Study** in *Annals of Surgery Open.*



Original Study

ANNALS OF SURGERY OPEN

OPEN

Mismatch Repair Status and Surgical Outcomes in Localized Colorectal Cancer

A Nationwide Cohort Study

Tobias Freyberg Justesen, MD,* Adile Orhan, MD,* Andreas Weinberger Rosen, MD,* Mikail Gögenur, MD,* and Ismail Gögenur, MD, DMSc*†

Objective: This study examined the association between deficient mismatch repair (dMMR) versus proficient MMR (pMMR) status and overall survival and disease-free survival in patients with localized colorectal cancer.
Background: Several distinctions exist between patients with dMMR and pMMR colorectal cancer. However, the impact on prognosis is yet to be investigated in large-scale cohort studies.
Methods: In this cohort study, we included patients who underwent curative-intent surgery for localized colorectal cancer between 2009 and 2020. Patients were identified in the Danish Colorectal Cancer Group database and patient-level data were extracted from 6 registry databases. After inclusion, patients with dMMR status were matched 1:1 to patients with pMMR status using an estimated propensity score.
Results: After matching, 5994 patients were included. The patients had a median age of 74 years and a median follow-up of 4.1 years. There was no significant association between mismatch repair (MMR) status and overall survival (hazard ratio, 0.91; 95% confidence interval [CI], 0.81–1.03) or disease-free survival (hazard ratio, 0.89; 95% CI, 0.78–1.01). However, the restricted 5-year mean disease-free survival time, calculated due to violation of the proportional hazards assumption, showed a significant absolute difference of 0.13 years (95% CI, 0.03–0.23; $P = 0.01$) in favor of the dMMR group.
Conclusions: No significant association with overall survival was found according to MMR status. dMMR status was, however, found to be associated with marginally improved disease-free survival compared to pMMR status in patients with localized colorectal cancer undergoing curative-intent surgery.

Keywords: colorectal cancer, deficient mismatch repair system, proficient mismatch repair system

OHDSI Shoutouts!

Congratulations to the team of **Rowdy de Groot, Frank van der Graaff, Daniël van der Doelen, Michiel Luijten, Ronald De Meyer, Hekmat Alrouh, Hedy van Oers, Jacintha Tieskens, Josjan Zijlmans, Meike Bartels, Arne Popma, Nicolette de Keizer, Ronald Cornet, and Tinca Polderman** on the publication of **Implementing Findable, Accessible, Interoperable, Reusable (FAIR) Principles in Child and Adolescent Mental Health Research: Mixed Methods Approach** in *JMIR Mental Health.*

Original Paper

Implementing Findable, Accessible, Interoperable, Reusable (FAIR) Principles in Child and Adolescent Mental Health Research: Mixed Methods Approach

Rowdy de Groot[1,2], MSc; Frank van der Graaff[3]; Daniël van der Doelen[4], MSc; Michiel Luijten[5,6,7], PhD; Ronald De Meyer[3], PhD; Hekmat Alrouh[8], MSc; Hedy van Oers[5,7], PhD; Jacintha Tieskens[9], PhD; Josjan Zijlmans[7,10], PhD; Meike Bartels[7,8], PhD; Arne Popma[7,10,11], PhD; Nicolette de Keizer[1,2], PhD; Ronald Cornet[1,2*], PhD; Tinca J C Polderman[4,7,9,10,11,12*], PhD

[1]Department of Medical Informatics, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, Netherlands
[2]Amsterdam Public Health, Digital Health, Amsterdam University Medical Center, Amsterdam, Netherlands
[3]Praktikon, Nijmegen, Netherlands
[4]Karakter Child and Adolescent Psychiatry University Centre, Nijmegen, Netherlands
[5]Emma Children's Hospital, Child and Adolescent Psychiatry & Psychosocial Care, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, Netherlands
[6]Amsterdam University Medical Center, Vrije Universiteit Amsterdam, Epidemiology and Data Science, Amsterdam, Netherlands
[7]Amsterdam Public Health, Mental Health, Amsterdam University Medical Center, Amsterdam, Netherlands
[8]Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[9]Leiden University Medical Center Curium–Child and Adolescent Psychiatry, Leiden University Medical Center, Leiden, Netherlands
[10]Department of Child and Adolescent Psychiatry & Psychosocial Care, Amsterdam University Medical Center, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[11]Levvel, Academic Center for Child and Adolescent Psychiatry, Amsterdam, Netherlands
[12]Department of Child and Adolescent Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, Netherlands
*these authors contributed equally

Corresponding Author:

Rowdy de Groot, MSc
Department of Medical Informatics
Amsterdam University Medical Center, University of Amsterdam
Meibergdreef 9
Amsterdam, 1105 AZ
Netherlands
Phone: 31 648499049
Email: rowdy.degroot@amsterdamumc.nl

Abstract

Background: The FAIR (Findable, Accessible, Interoperable, Reusable) data principles are a guideline to improve the reusability of data. However, properly implementing these principles is challenging due to a wide range of barriers.

Objectives: To further the field of FAIR data, this study aimed to systematically identify barriers regarding implementing the FAIR principles in the area of child and adolescent mental health research, define the most challenging barriers, and provide recommendations for these barriers.

OHDSI Shoutouts!

Congratulations to the team of **Chen Yanover, Ramit Magen-Rimon, Erica A. Voss, Joel Swerdel, Anna Sheahan, Nathan Hall, Jimyung Park, Rae Woong Park, Kwang Jae Lee, Sung Jae Shin, Seung In Seo, Kyung-Joo Lee, Thomas Falconer, Leonard Haas, Paul Nagy, Mary Grace Bowring, Michael Cook, Steven Miller, Tal El-Hay, Maytal Bivas-Benita, Pinchas Akiva, Yehuda Chowers and Roni Weisshof** on the publication of **Characteristics and Outcomes of Over a Million Patients with Inflammatory Bowel Disease in Seven Countries: Multinational Cohort Study and Open Data Resource** in *Digestive Diseases & Sciences.*

Digestive Diseases and Sciences
https://doi.org/10.1007/s10620-024-08787-x

ORIGINAL ARTICLE

Check for updates

Characteristics and Outcomes of Over a Million Patients with Inflammatory Bowel Disease in Seven Countries: Multinational Cohort Study and Open Data Resource

Chen Yanover[1,2] · Ramit Magen-Rimon[1,3] · Erica A. Voss[1,4] · Joel Swerdel[1,4] · Anna Sheahan[1,4] · Nathan Hall[1,4] · Jimyung Park[1,5,6] · Rae Woong Park[1,6] · Kwang Jae Lee[1,7] · Sung Jae Shin[1,7] · Seung In Seo[1,8] · Kyung-Joo Lee[1,9] · Thomas Falconer[1,5] · Leonard Haas[1,10] · Paul Nagy[1,10] · Mary Grace Bowring[1,10] · Michael Cook[1,10] · Steven Miller[1,10] · Tal El-Hay[1,2] · Maytal Bivas-Benita[1,2] · Pinchas Akiva[1,2] · Yehuda Chowers[1,11] · Roni Weisshof[1,11]

Received: 10 September 2024 / Accepted: 4 December 2024

Abstract

Background and Aims Observational healthcare data are an important tool for delineating patients' inflammatory bowel disease (IBD) journey in real-world settings. However, studies that characterize IBD cohorts typically rely on a single resource, apply diverse eligibility criteria, and extract variable sets of attributes, making comparison between cohorts challenging. We aim to longitudinally describe and compare IBD patient cohorts across multiple geographic regions, employing unified data and analysis framework.

Methods We conducted a descriptive cohort study, using routinely collected healthcare data, from a federated network of data partners in sixteen databases from seven countries (USA, UK, France, Germany, Japan, Korea, and Australia); and computed the prevalence of thousands of attributes, across multiple baseline and follow-up time windows, for full disease cohorts and various strata.

Results Characterizing the disease trajectory of 462,502 Crohn's disease (CD) and 589,118 ulcerative colitis (UC) subjects, we observed a decline over time in the average age at CD diagnosis in Europe and North America but less pronounced shifts in Japan and Korea; an uptick in the proportion of patients with anxiety diagnosis prior to CD diagnosis in European and US datasets; and stable rates of segmental colonic and small bowel resections within one and three years following UC and CD diagnosis, respectively, in most US databases.

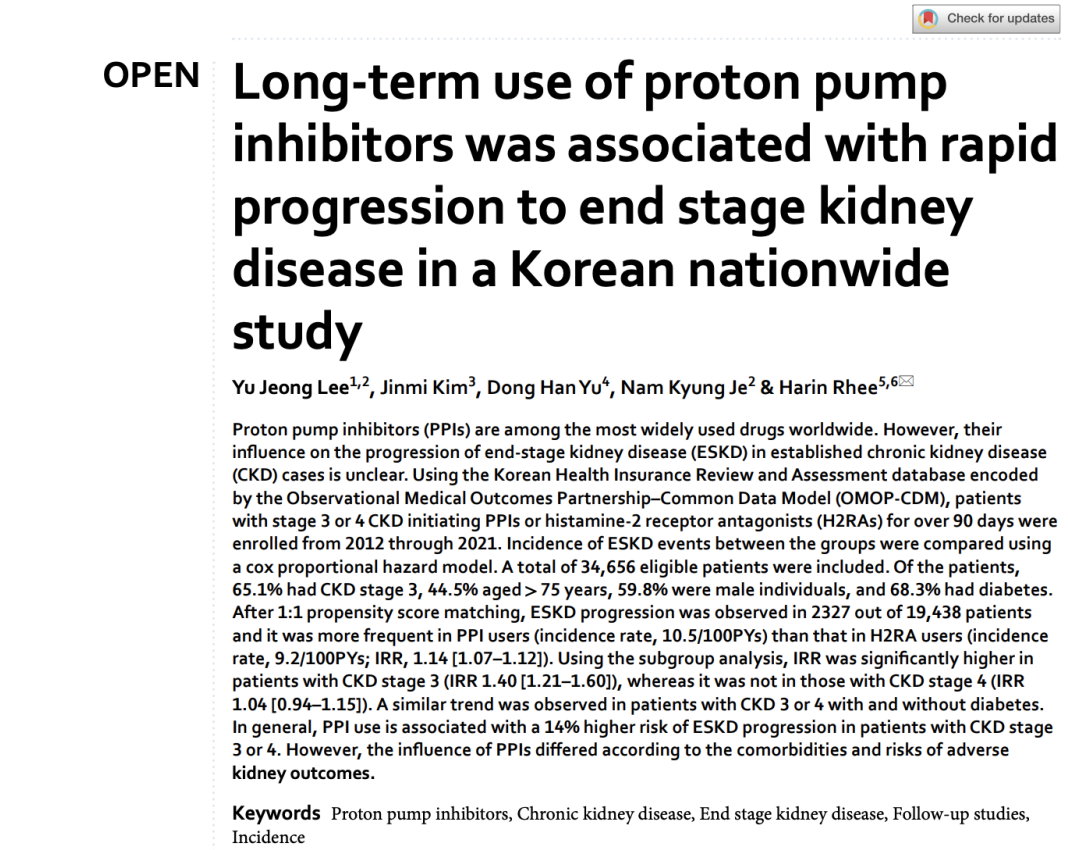
Conclusions The study provides a comprehensive characterization of IBD patient cohorts from various countries including insights into disease trends, demographics, and pre-diagnosis symptoms. All characteristics and outcomes are publicly available, providing an unprecedented, comprehensive open resource for clinicians and researchers.

Keywords Crohn's disease · Ulcerative colitis · Routinely collected health data · Cohort study

OHDSI Shoutouts!

Congratulations to the team of **Yu Jeong Lee, Jinmi Kim, Dong Han Yu, Nam Kyung Je and Harin Rhee** on the publication of **Long-term use of proton pump inhibitors was associated with rapid progression to end stage kidney disease in a Korean nationwide study** in *Scientific Reports*.

scientific reports



Check for updates

OPEN

Long-term use of proton pump inhibitors was associated with rapid progression to end stage kidney disease in a Korean nationwide study

Yu Jeong Lee[1,2], Jinmi Kim[3], Dong Han Yu[4], Nam Kyung Je[2] & Harin Rhee[5,6]✉

Proton pump inhibitors (PPIs) are among the most widely used drugs worldwide. However, their influence on the progression of end-stage kidney disease (ESKD) in established chronic kidney disease (CKD) cases is unclear. Using the Korean Health Insurance Review and Assessment database encoded by the Observational Medical Outcomes Partnership–Common Data Model (OMOP-CDM), patients with stage 3 or 4 CKD initiating PPIs or histamine-2 receptor antagonists (H2RAs) for over 90 days were enrolled from 2012 through 2021. Incidence of ESKD events between the groups were compared using a cox proportional hazard model. A total of 34,656 eligible patients were included. Of the patients, 65.1% had CKD stage 3, 44.5% aged > 75 years, 59.8% were male individuals, and 68.3% had diabetes. After 1:1 propensity score matching, ESKD progression was observed in 2327 out of 19,438 patients and it was more frequent in PPI users (incidence rate, 10.5/100PYs) than that in H2RA users (incidence rate, 9.2/100PYs; IRR, 1.14 [1.07–1.12]). Using the subgroup analysis, IRR was significantly higher in patients with CKD stage 3 (IRR 1.40 [1.21–1.60]), whereas it was not in those with CKD stage 4 (IRR 1.04 [0.94–1.15]). A similar trend was observed in patients with CKD 3 or 4 with and without diabetes. In general, PPI use is associated with a 14% higher risk of ESKD progression in patients with CKD stage 3 or 4. However, the influence of PPIs differed according to the comorbidities and risks of adverse kidney outcomes.

Keywords Proton pump inhibitors, Chronic kidney disease, End stage kidney disease, Follow-up studies, Incidence

OHDSI Shoutouts!

Congratulations to the team of **Harry-Anton Talvik, Marek Oja, Sirli Tamm, Kerli Mooses, Dage Särg, Marcus Lõo, Õie Renata Siimon, Hendrik Šuvalov, Raivo Kolde, Jaak Vilo, Sulev Reisberg, and Sven Laur** on the publication of **Repeatable process for extracting health data from HL7 CDA documents** in the *Journal of Biomedical Informatics*.

Journal of Biomedical Informatics 161 (2025) 104765

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Repeatable process for extracting health data from HL7 CDA documents

Harry-Anton Talvik [a,b], Marek Oja [a], Sirli Tamm [a], Kerli Mooses [a,*], Dage Särg [a,c], Marcus Lõo [a,d], Õie Renata Siimon [a], Hendrik Šuvalov [a], Raivo Kolde [a], Jaak Vilo [a,b], Sulev Reisberg [a,b], Sven Laur [a,b]

[a] Institute of Computer Science, University of Tartu, 51009 Tartu, Estonia
[b] STACC, 51009 Tartu, Estonia
[c] Estonian Genome Centre, Institute of Genomics, University of Tartu, 51010 Tartu, Estonia
[d] Institute of Mathematics and Statistics, University of Tartu, 51009 Tartu, Estonia

ARTICLE INFO

Keywords:
HL7 Clinical Document Architecture
ETL
Workflow
Pipeline
OMOP CDM
NLP

ABSTRACT

Objective: This study aims to address the gap in the literature on converting real-world Clinical Document Architecture (CDA) data into the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), focusing on the initial steps preceding the mapping phase. We highlight the importance of a repeatable Extract-Transform-Load (ETL) pipeline for health data extraction from HL7 CDA documents in Estonia for research purposes.

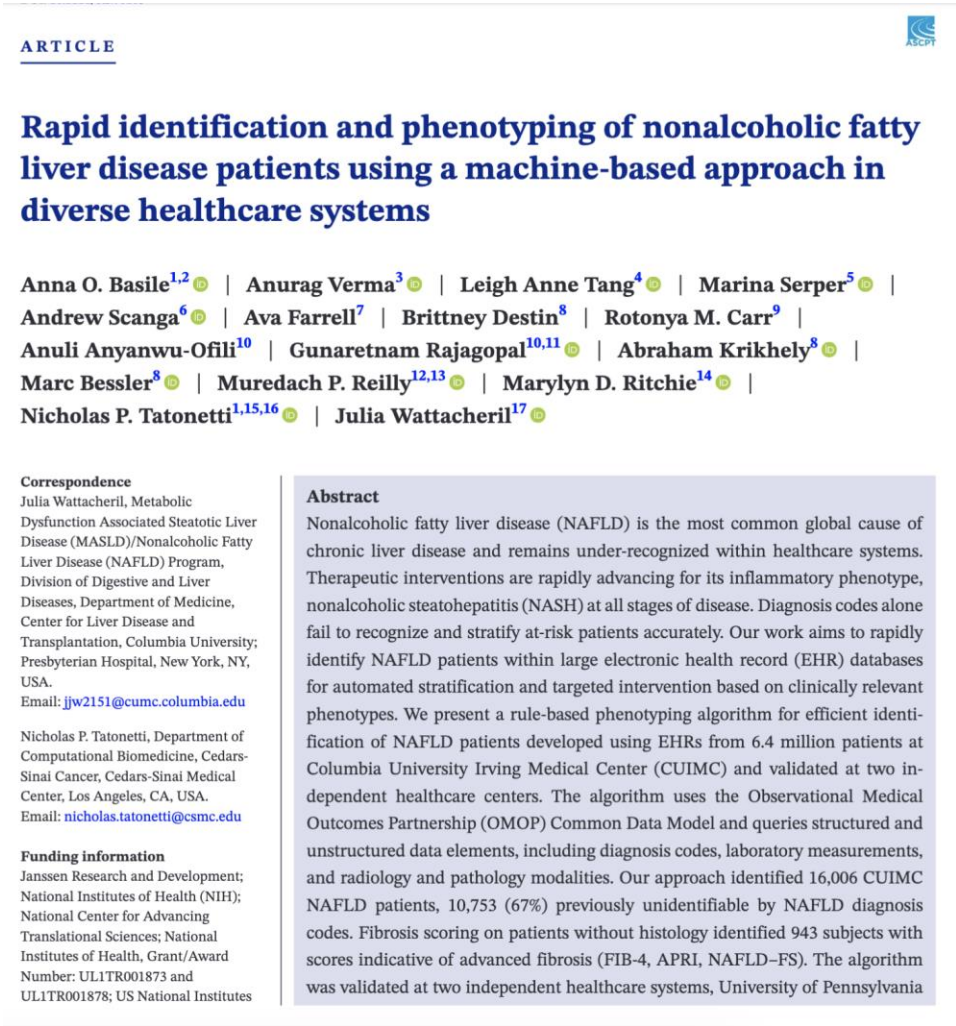
Methods: We developed a repeatable ETL pipeline to facilitate the extraction, cleaning, and restructuring of health data from CDA documents to OMOP CDM, ensuring a high-quality and structured data format. This pipeline was designed to adapt to continuously updated data exchange format changes and handle various CDA document subsets for different scientific studies.

Results: We demonstrated via selected use cases that our pipeline successfully transformed a significant portion of diagnosis codes, body weight and eGFR measurements, and PAP test results from CDA documents into OMOP CDM, showing the ease of extracting structured data. However, challenges such as harmonising diverse coding systems and extracting lab results from free-text sections were encountered. The iterative development of the pipeline facilitated swift error detection and correction, enhancing the process's efficiency.

Conclusion: After a decade of focused work, our research has led to the development of an ETL pipeline that effectively transforms HL7 CDA documents into OMOP CDM in Estonia, addressing key data extraction and transformation challenges. The pipeline's repeatability and adaptability to various data subsets make it a valuable resource for researchers dealing with health data. While tested on Estonian data, the principles outlined are broadly applicable, potentially aiding in handling health data standards that vary by country. Despite newer health data standards emerging, the relevance of CDA for retrospective health studies ensures the continuing importance of this work.

OHDSI Shoutouts!

Congratulations to the team of **Anna O. Basile, Anurag Verma, Leigh Anne Tang, Marina Serper, Andrew Scanga, Ava Farrell, Brittney Destin, Rotonya M. Carr, Anuli Anyanwu-Ofili, Gunaretnam Rajagopal, Abraham Krikhely, Marc Bessler, Muredach P. Reilly, Marylyn D. Ritchie, Nicholas P. Tatonetti, and Julia Wattacheril** on the publication of **Rapid identification and phenotyping of nonalcoholic fatty liver disease patients using a machine-based approach in diverse healthcare systems** in *Clinical and Translational Sciences*.



ARTICLE

Rapid identification and phenotyping of nonalcoholic fatty liver disease patients using a machine-based approach in diverse healthcare systems

Anna O. Basile[1,2] | Anurag Verma[3] | Leigh Anne Tang[4] | Marina Serper[5] | Andrew Scanga[6] | Ava Farrell[7] | Brittney Destin[8] | Rotonya M. Carr[9] | Anuli Anyanwu-Ofili[10] | Gunaretnam Rajagopal[10,11] | Abraham Krikhely[8] | Marc Bessler[8] | Muredach P. Reilly[12,13] | Marylyn D. Ritchie[14] | Nicholas P. Tatonetti[1,15,16] | Julia Wattacheril[17]

Correspondence
Julia Wattacheril, Metabolic Dysfunction Associated Steatotic Liver Disease (MASLD)/Nonalcoholic Fatty Liver Disease (NAFLD) Program, Division of Digestive and Liver Diseases, Department of Medicine, Center for Liver Disease and Transplantation, Columbia University; Presbyterian Hospital, New York, NY, USA.
Email: jjw2151@cumc.columbia.edu

Nicholas P. Tatonetti, Department of Computational Biomedicine, Cedars-Sinai Cancer, Cedars-Sinai Medical Center, Los Angeles, CA, USA.
Email: nicholas.tatonetti@csmc.edu

Funding information
Janssen Research and Development; National Institutes of Health (NIH); National Center for Advancing Translational Sciences; National Institutes of Health, Grant/Award Number: UL1TR001873 and UL1TR001878; US National Institutes

Abstract
Nonalcoholic fatty liver disease (NAFLD) is the most common global cause of chronic liver disease and remains under-recognized within healthcare systems. Therapeutic interventions are rapidly advancing for its inflammatory phenotype, nonalcoholic steatohepatitis (NASH) at all stages of disease. Diagnosis codes alone fail to recognize and stratify at-risk patients accurately. Our work aims to rapidly identify NAFLD patients within large electronic health record (EHR) databases for automated stratification and targeted intervention based on clinically relevant phenotypes. We present a rule-based phenotyping algorithm for efficient identification of NAFLD patients developed using EHRs from 6.4 million patients at Columbia University Irving Medical Center (CUIMC) and validated at two independent healthcare centers. The algorithm uses the Observational Medical Outcomes Partnership (OMOP) Common Data Model and queries structured and unstructured data elements, including diagnosis codes, laboratory measurements, and radiology and pathology modalities. Our approach identified 16,006 CUIMC NAFLD patients, 10,753 (67%) previously unidentifiable by NAFLD diagnosis codes. Fibrosis scoring on patients without histology identified 943 subjects with scores indicative of advanced fibrosis (FIB-4, APRI, NAFLD–FS). The algorithm was validated at two independent healthcare systems, University of Pennsylvania

OHDSI Shoutouts!

Congratulations to the team of **Rowdy de Groot, Savannah Glaser, Alexandra Kogan, Stephanie Medlock, Anna Alloni, Matteo Gabetta, Szymon Wilk, Nicolette de Keizer, and Ronald Cornet** on the publication of **ATC-to-RxNorm mappings - A comparison between OHDSI Standardized Vocabularies and UMLS Metathesaurus** in the *International Journal of Medical Informatics*.

International Journal of Medical Informatics 195 (2025) 105777

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

ATC-to-RxNorm mappings – A comparison between OHDSI Standardized Vocabularies and UMLS Metathesaurus

Rowdy de Groot [a,b,*], Savannah Glaser [a,b], Alexandra Kogan [c], Stephanie Medlock [a,b], Anna Alloni [d], Matteo Gabetta [d], Szymon Wilk [e], Nicolette de Keizer [a,b], Ronald Cornet [a,b]

[a] Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Amsterdam, the Netherlands
[b] Methodology, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands
[c] Department of Information Systems, University of Haifa, Haifa, Israel
[d] Biomeris S.r.l., Pavia, Italy
[e] Institute of Computing Science, Poznan University of Technology, Poznan, Poland

ARTICLE INFO

Keywords:
Terminology mapping
Data interoperability
ATC
RxNorm
OHDSI Standardized Vocabularies
UMLS

ABSTRACT

Introduction: The World Health Organization global standard for representing drug data is the Anatomical Therapeutic Chemical (ATC) classification. However, it does not represent ingredients and other drug properties required by clinical decision support systems. A mapping to a terminology system that contains this information, like RxNorm, may help fill this gap. This work evaluates and compares the completeness of mappings from the chemical substance level (5th-level) ATC classes to RxNorm ingredient concepts in the OHDSI Standardized Vocabularies (OSV) and the Unified Medical Language System (UMLS) Metathesaurus.

Methods: To check the concordance between OSV and UMLS we compared the included contents of ATC and RxNorm not only in OSV and UMLS but also in BioPortal and the National Library of Medicine (NLM) repository. For each repository, we determined the number of 5th-level ATC concepts, RxNorm ingredient concepts, missing classes and concepts, and the ATC categories with the most missing concepts. The mappings from ATC to RxNorm in OSV and UMLS were compared, and we determined the number of mappings in common, and the mapping differences, which we categorized. We applied the mappings from OSV and UMLS on a sample of Electronic Health Record (EHR) data.

Results: NLM contained the most ATC and RxNorm concepts. UMLS contained more missing mappings (null mappings) than OSV, 1949 versus 916. Most mapping differences were in the "unknown ingredient in the ATC label" category, for which UMLS provided no mappings. UMLS had a higher coverage of mappings in the sample EHR data than OSV, 96.5% versus 91%.

Discussion: In conclusion, opting for OSV rather than UMLS is generally preferable for an ATC to RxNorm mapping since OSV provides more mappings. However, the results of the sample data show that UMLS can have fewer null mappings in concrete applications.

Three Stages of The Journey

Where Have We Been?

Where Are We Now?

Where Are We Going?

Upcoming Workgroup Calls

Date	Time (ET)	Meeting
Tuesday	12 pm	Common Data Model Vocabulary Subgroup
Wednesday	9 am	Patient-Level Prediction
Wednesday	12 pm	Health Equity
Wednesday	2 pm	Natural Language Processing
Wednesday	7 pm	Medical Imaging
Thursday	10:30 am	Evidence Network
Thursday	12 pm	Strategus HADES Subgroup
Thursday	6 pm	Eyecare and Vision Research
Friday	9 am	Phenotype Development and Evaluation
Friday	11:30 am	Steering
Friday	11 pm	China Chapter
Monday	9 am	Vaccine Vocabulary
Monday	10 am	Healthcare Systems Interest Group

OHDSI Europe Symposium - Save-the-date!

OHDSI BELGIUM

EUROPE

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS

Save-the-date

5-7 July 2025

Location

Old Prison - Hasselt University
Martelarenlaan
Hasselt - BELGIUM

Next CBER Best Seminar: Jan. 15

Topic: Emulation of Target Trial on Vaccinations During Pregnancy

Presenter: Sonia Hernández-Díaz, MD, DrPH, Professor of Epidemiology, Harvard T.H. Chan School of Public Health

Date/Time: Jan. 15, 11 am ET

ohdsi.org/cber-best-seminar-series

Collaborator Spotlight: Linying Zhang

Linying Zhang is an assistant professor of
at Washington University School of Medicine in St. Louis. She earned the 2024 Titan Award for Methodological Research.

In the latest edition of the Collaborator Spotlight, Linying discusses OHDSI impact in her PhD journey, the critical need for methods research innovations, bringing Washington University into the OHDSI Evidence Network and … worm composting.

ohdsi.org/spotlight-linying-zhang

January Newsletter

OHDSI

The Journey Newsletter (January 2025)

Happy New Year! OHDSI spent 2024 advancing its mission of improving health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care. We reflect on some of what made 2024 special—and we look forward to the possibilities for 2025—in the latest edition of The Journey newsletter! **#JoinTheJourney**

Video Podcast: Looking Back On 2024 Progress

In the latest On The Journey videocast, Patrick Ryan and Craig Sachson reflect on progress made within OHDSI in the areas of open-science collaboration, community data standards, open-source development and evidence generation before looking ahead to setting the course for 2025. *(if video does not appear, please click view this email in your browser)*

Community Updates

Where Have We Been?

• The 2024 Asia-Pacific (APAC) Symposium was held Dec. 4-8 in Singapore, and it included a day of tutorials and a two-day main conference. Slides and photos can be found on the event homepage. Thank you to Mengling 'Mornin' Feng and his team for leading this symposium.
• December concluded a record-breaking year for published studies relating to OMOP or OHDSI tool/practices. Including the publications listed below, there were more than 135 studies shared in peer-reviewed journals this year.

Where Are We Now?

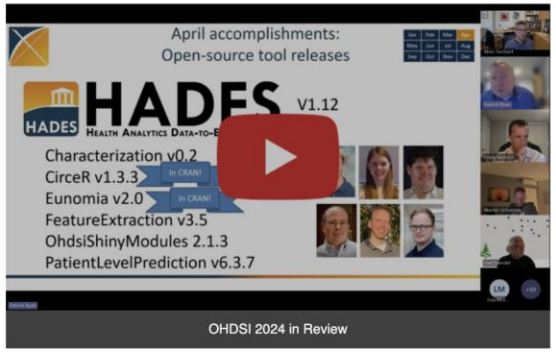
• The first community call of 2025 will be Tuesday, Jan. 7, at 11 am ET. Patrick Ryan will lead a session that focuses on where OHDSI can go over the next 12 months. A new link for the call has been sent out, and it will be available on the community calls page.
• The first CBER BEST seminar of 2025 will be held Jan. 15 and will be led by Harvard's Sonia Hernández-Díaz. More information and the call link is available in this newsletter.
• Workgroup leads should focus on helping set their objectives and key results (OKRs) for 2025. These will be shared throughout February community calls.

Where Are We Going?

• The 2025 Europe Symposium will be held July 5-7 at the historic "Old Prison" at UHasselt in Belgium. More details will be announced when available.
• The #OHDSISocialShowcase features posters, software demos and lightning talks from the 2024 Global Symposiums this month. Please make sure you are following OHDSI's LinkedIn, Twitter/X and Instagram feeds to receive daily updates on the research presented by our community.

Review: OHDSI in 2024

During the Dec. 10 community call, Patrick Ryan reflected on the goals, accomplishments and unfinished work for the OHDSI community in 2024. The video presentation, which is available below, highlights monthly studies, open-source developments, community activities, individual accomplishments, and plenty more.



OHDSI 2024 in Review

Preview: OHDSI in 2025

So what comes next for the OHDSI community? We will start to set the path for OHDSI work in 2025 during our opening community call of the year on Tuesday, Jan. 7, at 11 am ET. A new call link went out already, and the link will also be available on our community calls page.

Community Calls Page

December Publications

Kang M, Alvarado-Guzman JA, Rasmussen LV, Starren JB. Evolution of a Graph Model for the OMOP Common Data Model. Appl Clin Inform. 2024 Oct;15(5):1056-1065. doi: 10.1055/s-0044-1791487. Epub 2024 Dec 4. PMID: 39631779; PMCID: PMC11617070.

Jeon K, Park WY, Kahn CE Jr, Nagy P, You SC, Yoon SH. Advancing Medical Imaging Research Through Standardization: The Path to Rapid Development, Rigorous Validation, and Robust Reproducibility. Invest Radiol. 2025 Jan 1;60(1):1-10. doi: 10.1097/RLI.0000000000001106. Epub 2024 Jul 11. PMID: 38985896.

Tran TC, Schlueter DJ, Zeng C, Mo H, Carroll RJ, Denny JC. PheWAS analysis on large-scale biobank data with PheTK. Bioinformatics. 2024 Dec 9:btae719. doi: 10.1093/bioinformatics/btae719. Epub ahead of print. PMID: 39657951.

Elhussein A, Baymuradov U; NYGC ALS Consortium; Elhadad N, Natarajan K, Gürsoy G. A framework for sharing of clinical and genetic data for precision medicine applications. Nat Med. 2024 Dec;30(12):3578-3589. doi: 10.1038/s41591-024-03239-5. Epub 2024 Sep 3. PMID: 39227443; PMCID: PMC11645287.

Zhang S, Cornet R, Benis N. Cross-Standard Health Data Harmonization using Semantics of Data Elements. Sci Data. 2024 Dec 19;11(1):1407. doi: 10.1038/s41597-024-04168-1. PMID: 39702578.

Justesen TF, Orhan A, Rosen AW, Gögenur M, Gögenur I. Mismatch Repair Status and Surgical Outcomes in Localized Colorectal Cancer: A Nationwide Cohort Study. Ann Surg Open. 2024 Oct 16;5(4):e499. doi: 10.1097/AS9.0000000000000499. PMID: 39711680; PMCID: PMC11661751.

de Groot R, van der Graaff F, van der Doelen D, Luijten M, De Meyer R, Alrouh H, van Oers H, Tieskens J, Zijlmans J, Bartels M, Popma A, de Keizer N, Cornet R, Polderman TJC. Implementing Findable, Accessible, Interoperable, Reusable (FAIR) Principles in Child and Adolescent Mental Health Research: Mixed Methods Approach. JMIR Ment Health. 2024 Dec 19;11:e59113. doi: 10.2196/59113. PMID: 39727091; PMCID: PMC11683739.

Yanover C, Magen-Rimon R, Voss EA, Swerdel J, Sheahan A, Hall N, Park J, Park RW, Lee KJ, Shin SJ, Seo SI, Lee KJ, Falconer T, Haas L, Nagy P, Bowring MG, Cook M, Miller S, El-Hay T, Bivas-Benita M, Akiva P, Chowers Y, Weisshof R. Characteristics and Outcomes of Over a Million Patients with Inflammatory Bowel Disease in Seven Countries: Multinational Cohort Study and Open Data Resource. Dig Dis Sci. 2024 Dec 26. doi: 10.1007/s10620-024-08787-x. Epub ahead of print. PMID: 39724470.

January Newsletter

CDM Survey Subgroup Landscape Assessment

The CDM Survey Subgroup invites colleagues who have or are going to design, develop, and/or implement research surveys and use them with the OMOP CDM to share information about those efforts by completing this survey. Your completion of this 10-15 minute survey will provide information to the CDM workgroup about OMOP utilization among survey research teams. The CDM Survey subgroup is a collaborative effort, led by a team at the National Cancer Institute, to develop standardized approaches and best practices for helping research teams better integrate survey data elements into the OMOP common data model.

The deadline has been extended to mid-January.

LANDSCAPE ASSESSMENT

- **Activities**
  - Invite representatives from cohorts with experience using the CDM for survey data to share their knowledge and challenges.
  - Conduct a community survey to gather information on experiences and needs related to survey data in the CDM.
  - Review the most used Common Data Elements (CDMs) as a foundation for developing standards, tools, and best practices.
- **Key Result**
  - A comprehensive report summarizing survey CDM mapping resources, challenges, and identified development priorities (vocabulary, standards, tools, best practices) to be shared with the OHDSI community.

WHO SHOULD PARTICIPATE

- You have survey data and you've mapped it to the OMOP CDM
- You have survey data and you would like to map it to the OMOP CDM
- You are in the process of developing a survey(s) and plan to map to the OMOP CDM
- Multiple perspectives from the same team
- Multiple surveys from the same person

#OHDSISocialShowcase This Week

Monday

Institutionalizing data interoperability and the application of common data models in a health data and research center: CIDACS' experience in Brazil

(Valentina Martufi, Juliana Araújo Prata de Faria, Danilo Luis Cerqueira Dias, Elzo Pereira Pinto Junior, Roberto Carreiro, Pablo Ivan Ramos, Maurício L. Barreto)

Institutionalizing data interoperability and the application of common data models in a health data and research center: CIDACS' experience in Brazil

Juliana Araújo Prata de Faria

INTRO

❖CIDACS' mission is to **expand the understanding of socio-economic and environmental determinants of health** and provide robust scientific evidence to support decision-makers, through:
- ✓ application of **innovative methodologies**;
- ✓ linkage of large **real-world data sets**;
- ✓ creation of cohorts including millions of individuals;

❖In January 2024 CIDACS instituted the **IDAF (Data Interoperability and Federated Analysis)**, as a new subgroup within its Data Platform team;
→ an initiative that recognized the **increasing global valorization** of the application of the **OMOP CDM** for the **generation of valuable and robust knowledge** in the field of health

METHODS

The objectives of the IDAF encompass:
1- Consolidating and adopting **open standards for health data** (OMOP CDM Model)
2- Supporting the **development of research projects** utilizing common data models
3 - Establishing a solid foundation to **promote data reuse** (according to the FAIR principles)
4 - **Consolidating interoperability processes** and federated data analyses
5 - **Training and disseminating** information on interoperability and common data models.

RESULTS
- ✓ Thanks to the OHDSI tools, the OMOP CDM will be applied to **CIDACS' cohorts**, as well as the other real-world data entrusted to CIDACS by the Brazilian government;
- →answer a vast **variety of research questions** related to the health of the global population;
- ✓ **Current efforts** are being focused towards **maternal and child health** outcomes;
- ✓ Nonetheless, it is envisioned that **CIDACS' OMOPped data** may provide the opportunity to delve into a **wide range of global health issues**, including **mental health**, the impact of **environmental factors and climate change** on the population's health as well as **health systems characteristics** and their influence on health outcomes.

CIDACS hopes to ***contribute to the OHDSI community*** the promotion of the inclusion of ***significant socioeconomic variables***, to provide a ***more holistic*** understanding of health determinants, boosting the application of the ***OMOP CDM in the Global South***.

Acknowledgments:
the creation of the IDAF was possible thanks to projects funded by the Rockerfeller Foundation and the Bill & Melinda Gates Foundation

Contact email:
juliana.faria@fiocruz.br

cidacs Centro de Integração de Dados e Conhecimentos para Saúde

FIOCRUZ | Bahia

Find out more about the IDAF

Find out more about CIDACS' 100m Brazilians cohort

WHY the IDAF???
→ **Acknowledging:**
- ✓ the **steep learning curve** for the utilization of the tools developed within the OHDSI community (unsustainable to be climbed anew by project leads and implementers, as new research projects appear), and
- ✓ the importance of **contributing to the OHDSI open-sources projects** (not feasible by project implementers on short timelines)

IDAF's functions include:
- ✓ mapping CIDACS' data to the OMOP CDM by using tools like White Rabbit and Rabbit in a Hat;
- ✓ implementing ETL (Extract, Transform, Load) processes;
- ✓ setting up OHDSI AWS (Amazon Web Services, Web API CDM by OHDSI), and configuring the Atlas tool for data analysis.
- ✓ Strengthening the OHDSI Brazilian Community through shared learning
- ✓ Contributing to the International OHDSI Ecosystem, especially in the development of new relevant tables and vocabularies (e.g. GIS)

Juliana Araújo Prata de Faria, Valentina Martufi, Danilo Luis Cerqueira Dias, Elzo Pereira Pinto Junior, Roberto Carreiro, Pablo Ivan Ramos, Maurício L. Barreto

UNIVERSITY NORTH STATE

OHDSI

#OHDSISocialShowcase This Week

Tuesday

Characterizing Phenotype Descriptions in All of Us Publications

(Emily Clark, Matthew Spotnitz, Lew Berman, John Giannini, Yechiam Ostchega, Lakshmi Priya Anandang)

OHDSI

Characterizing Phenotype Descriptions in *All of Us* Publications

All of Us — THE FUTURE OF HEALTH BEGINS WITH YOU

Emily Clark[1], MPH, Matthew Spotnitz[2], MD, MPH, John Giannini[2], PhD, Lakshmi Priya Anandan[3], MPH, Yechiam Ostchega[2], PhD, RN, Lew Berman[2], PhD, MS
[1] GAP Solutions, Inc., [2] National Institutes of Health / All of Us Research Program, [3]Leidos

Background

Phenotypes are an essential component of observational healthcare research, and the basis for a myriad of patient cohorts. **There is substantial variability in the methods used for defining and describing phenotypes.**

The *All of Us* research program is an effort to aggregate data from a diverse cohort of 1 million or more participants from across the United States. Data collected for the platform are from electronic health records (EHRs), biospecimens, surveys, and other sources, including systems that generate FHIR-compliant data. Currently over 10,000 researchers have registered to use this data. Data is accessible in a workbench and is formatted in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

Our aim was to evaluate the variability in phenotypic definitions in *All of Us* publications.

Methods

- We (E.C.) manually reviewed the list of published papers from the program inception in 2017 until May 15, 2023.
- Papers were included if authors studied a phenotype, and papers with multiple phenotypes were evaluated on the main phenotype.
- We excluded papers that described program operations or genome-wide association studies.
- This study only described which source coding systems were used.
- 101 out of 176 papers considered met the inclusion criteria.

Results

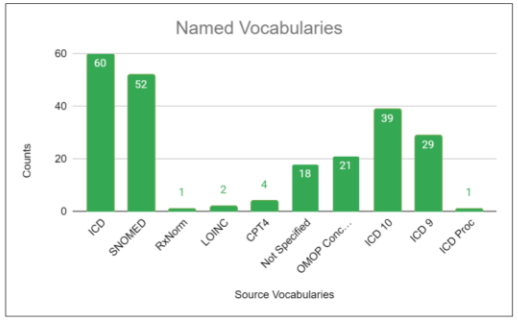


Figure 1. Bar chart of source codes from the 101 publications. Some papers described using more than one source vocabulary.

Each of these papers were reviewed manually, and of those, 101 (57%) were included in this analysis. Of the studies included:

- 60 (59%) phenotypes mentioned ICD diagnostic codes,
- 52 (51%) that mentioned SNOMED codes,
- 21 (21%) that mentioned OMOP concept IDs, and
- 18 (18%) papers that did not mention any data standard or set of source codes.

Discussion and Limitations

- Most papers concentrated on ICD and SNOMED codes rather than OMOP Concept IDs.
- This analysis was limited in that: 1) it did not compare the reported codes to the actual codes used in analysis, and 2) papers published after May 15, 2023 were not included in the analysis.
- Future research could consider: 1) explanations for the high frequency of ICD codes compared to SNOMED or OMOP concepts and diagnosis codes compared to data from other domains, or 2) methods for promoting consistent use of standardized vocabularies to improve the rigor of phenotyping.

Conclusions

Addressing the overutilization of ICD and SNOMED concepts compared to others, the relative underutilization of procedure codes, and setting standards for phenotype definitions **may improve the reliability and accuracy of observational health research.**

- Most phenotypes used ICD diagnosis codes, a non-standard source code, which was followed closely by the OMOP standardized vocabulary, SNOMED.
- Few studies described OMOP CDM concepts, despite that OMOP CDM is the data model for the *All of Us* Research Program.
- Diagnosis codes were used more than the other data domains, regardless of the source vocabulary.

References

1. *All of Us* data methods. National Institutes of Health *All of Us* Research Program. Accessed June 12, 2024. https://www.researchallofus.org/data-tools/methods/
2. Brandt PS, Kho A, Luo Y, et al. Characterizing variability of electronic health record-driven phenotype definitions. *J Am Med Inform Assoc*. 2023;30(3):427-437. doi:10.1093/jamia/ocac235
3. *All of Us* publications. National Institutes of Health *All of Us* Research Program. Accessed June 12, 2024. https://www.researchallofus.org/publications

Contact: Emily.clark3@nih.gov

#OHDSISocialShowcase This Week

Wednesday

Fine-Tuning Foundational AI Models to Code Diagnoses from Veterinary Health Records

(Mayla R. Boguslav, Adam Kiehl, Michael Kirby, David Kott, Nadia Saklou, G. Joseph Strecker, Terri Ward, Tracy Webb)

Fine-Tuning Foundational Models to Code Veterinary Diagnoses from Health Records

Adam Kiehl, Mayla R. Boguslav, David Kott, G. Joseph Strecker, Nadia Saklou, Tracy Webb, Terri Ward, Michael Kirby

INTRO

Goal: Make veterinary records more accessible in a research context
- *Extractable, shareable, usable*

Challenge: Free-text clinical notes difficult to mine for meaningful information
- Advanced NLP methods needed to label textual data

Example Record:

Diagnosis
- Sepsis
- Infectious vegetative endocarditis
- Bilateral hyphema
- Disseminated Intravascular Coagulation (DIC)

Assessment
Patient presented on [DATE] for severe lethargy that began the night before. [PATIENT] was ataxic and had a fever of 106 deg F at presentation to the VTH. [PATIENT] was exhibiting signs of sepsis...

Codes
238150007 (sepsis syndrome), 56819008 (endocarditis), 75229002 (hyphema), 67406007 (disseminated intravascular coagulation)

Inspiration: VetTag (2019) utilized CSU labeled data to train transformer model for smaller-scale diagnosis coding.

Innovation: Availability of pretrained LLMs expanded greatly since VetTag created, providing opportunity to leverage them.

METHODS

Data: 246,473 clinical notes labeled by CSU medical records team featuring 7,739 distinct SNOMED diagnosis codes.

Models: Ten pretrained LLMs downloaded from HuggingFace - human, veterinary, and general.

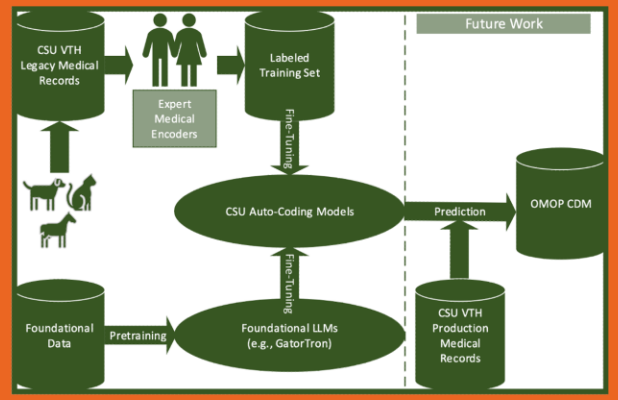
Augmentation: Dropout and classifier layers appended; fine-tuned as many transformer blocks as possible.

Fine-Tuning: LLMs fine-tuned for downstream multi-label classification task.

Computing: Training performed on HPC clusters of either Tesla V100 or A100 GPUs.

Metrics: Fine-tuned models evaluated using F1, precision, recall, and exact match rate.

We demonstrate the ability of fine-tuned LLMs to label clinical text with 7,739 SNOMED diagnosis codes as a proof of concept.



Key Takeaways:
1. Automated clinical coding is promising, but best as an *assistive* tool.
2. Task is not overly sensitive to size or domain of foundational LLM.
3. Similar results can still be achieved with limited fine-tuning data.

	Model	Parameters	F1	Precision	Recall	Exact Match	Fine-Tuning Time
Human Clinical LLMs	GatorTron	3.9B	**74.9**	**80.8**	**71.8**	**51.6**	21.9 hrs
	ClinicalBERT	135M	68.7	78.0	63.9	45.2	1.5 hrs
	medAlpaca	6.6B	67.1	79.1	61.7	41.6	14.1 hrs
Veterinary Clinical LLMs	VetBERT	108M	69.5	78.7	64.7	46.5	2.9 hrs
	PetBERT	108M	69.4	77.6	65.3	46.4	2.7 hrs
Non-Clinical LLMs	BERT Base	108M	68.5	77.5	63.9	45.8	3.6 hrs
	BERT Large	335M	70.4	78.4	66.2	47.2	7.4 hrs
	RoBERTa	125M	67.7	76.5	63.3	44.6	2.6 hrs
	GPT-2	124M	68.3	78.3	63.4	44.6	4.1 hrs
	GPT-2 XL	1.6B	71.7	80.4	67.2	47.8	15.2 hrs
	VetTag	*42M*	*66.2*	*72.1*	*63.1*	*26.2*	*Unknown*

RESULTS

Best Results: Fine-tuned GatorTron achieved average F1 of 74.9 and exact match rate of 51.6%.

Out-of-the-Box: Models performed poorly without fine-tuning on CSU labeled data.

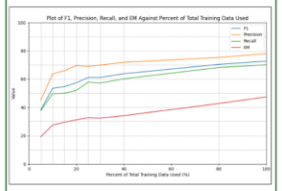
DISCUSSION

Best Results: Best results were obtained by fine-tuning GatorTron.

Model Size: Better results achieved using relatively large models, but comparable results observed with much smaller models.

Pretraining: Results were not sensitive to domain of LLM pretraining data.

Fine-Tuning: Results were not very sensitive to volume of fine-tuning data past a certain threshold.



Application: Could be used as assistive tool to help clinicians code diagnosis at time of entry - maintains human in the loop.

FUTURE WORK

While these results are encouraging, further study is needed to determine suitability for production uses.

Research Directions:
- Model explainability
- Generalizability
- Detailed error analysis
- Additional diagnosis text
- Synthetic data
- Ensemble models
- Additional LLMs

Data Governance: *Policies and procedures for data and model sharing will need to be developed to advance research aims.*

One Health Image: a4nh.cigar.org
VetTag: Zhang, Y., Nie, A., Zehnder, A., Page, R. L., & Zou, J. (2019). Vettag: Improving automated veterinary diagnosis coding via large-scale language modeling. NPJ digital medicine, 2 (1), 35.
Acknowledgements: A. Kiehl, G. Strecker, and T. Webb were supported by NIH/NCATS Colorado CTSA Grant Number UM1 TR004399

OHDSI

#OHDSISocialShowcase This Week

Thursday

Brain-penetrant calcium channel blockers for psychiatric use: revisiting the evidence for benefit

(David M Kern, Justin Bohn, Michael Maher, Dmytro Dymshyts, Azza Shoaibi)

Brain-penetrant calcium channel blockers for psychiatric use: Revisiting the evidence for benefit

PRESENTER: **Dave Kern**

BACKGROUND

- Multiple SNPs of a gene related to calcium channel gating (CACNA1C) have been linked to bipolar disorder and schizophrenia
- Recent RWE study showed that calcium channel blockers (CCBs) prescribed for hypertension may be protective against psychiatric disorders; however, there were concerns about the study design
- This study applied best practices in pharmacoepidemiology to better understand the relationship between CCBs and incidence of neuropsychiatric outcomes

METHODS

- Incidence of four outcomes were assessed: schizophrenia, schizoaffective disorder, major depressive disorder, bipolar disorder
- Large scale propensity scores were used to match users of brain-penetrant CCBs (BP-CCB) to those receiving non-brain-penetrant CCBs (NP-CCB).
- Intent to treat and on-treatment analyses were performed.
- Nine (9) real-world observational data sources were used from across the globe, including EHR and administrative claims data

Table 1. Patient counts before and after matching

Database Name	BP-CCBs	NP-CCBs	Matched
MarketScan Commercial Claims	215,449	1,741,739	172,598
MarketScan Medicaid	27,059	231,122	20,632
MarketScan Medicare	105,977	577,467	84,282
OPTUM Clinformatics	203,686	1,541,095	166,377
Optum EHR	325,655	3,240,419	275,079
PharMetrics	170,438	1,491,106	103,560
IQVIA Germany Disease Analyzer	44,956	238,703	22,443
IQVIA France Disease Analyzer	1,188	16,833	0
IQVIA Australian LPD	171	2,473	156
JMDC	60,741	268,817	36,631
Total	1,155,320	9,349,774	881,758

Matched is the number of individuals in *each cohort* after propensity score matching

Brain penetrant calcium channel blockers do not appear to reduce risk of neuropsychiatric outcomes when given at current therapeutic doses

RESULTS

Figure 1. Covariate balance before and after PS matching

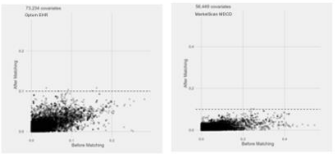
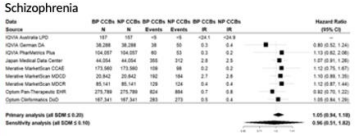
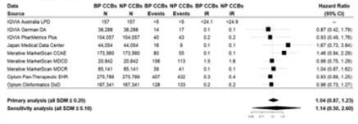


Figure 2. Incidence rates and forest plots of intent-to-treat analysis

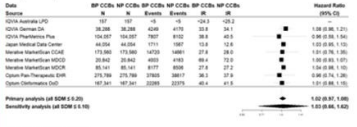
Schizophrenia



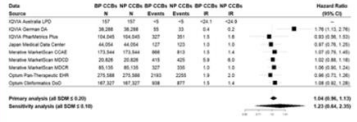
Schizoaffective disorder



Major depressive disorder



Bipolar disorder



David M. Kern, Justin Bohn, Michael Maher, Dmytro Dymshyts, Azza Shoaibi

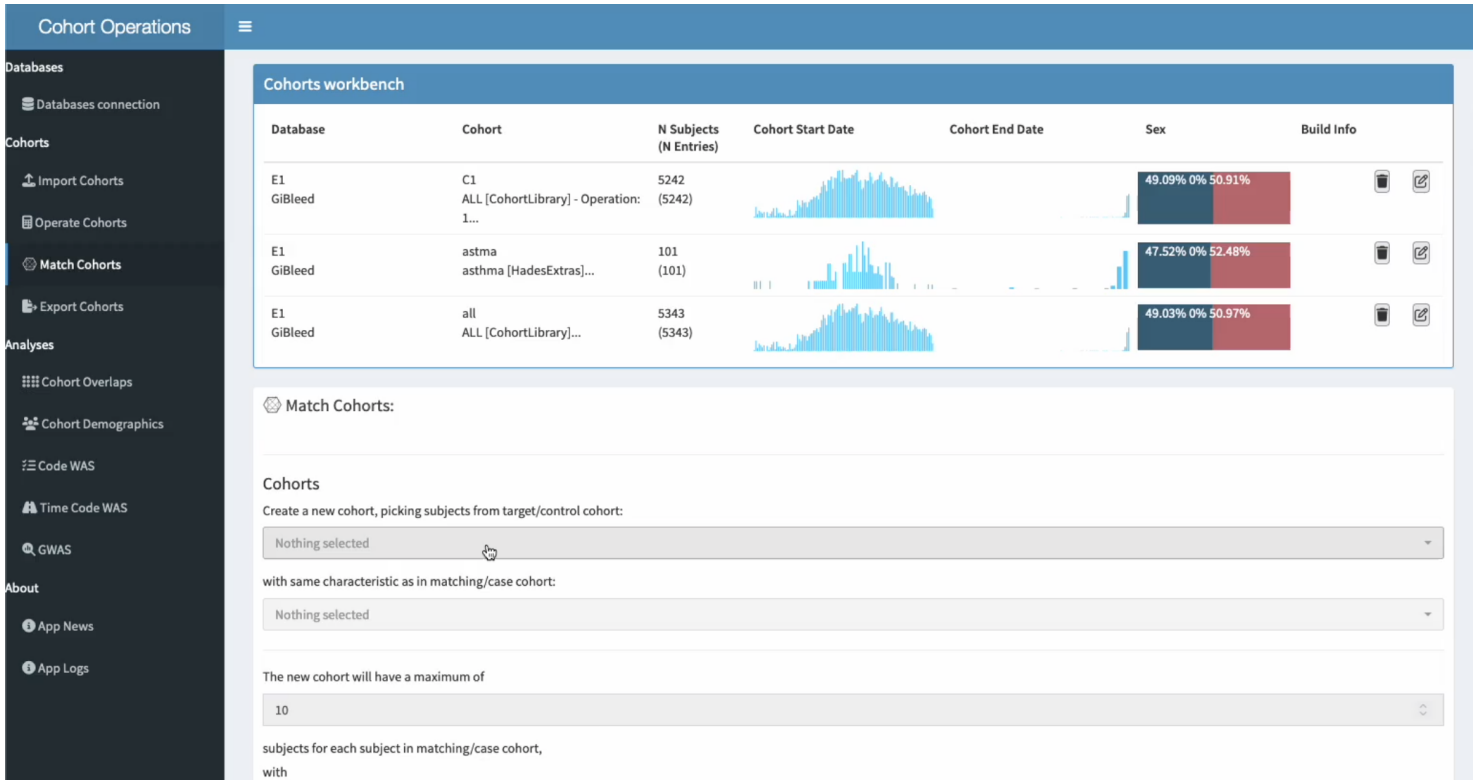
OHDSI

#OHDSISocialShowcase This Week

Friday

CohortOperations: A Modular Web Tool for Enhanced Cohort Analysis on the OMOP-CDM

(Javier Gracia-Tabuenca, Harri Siirtola, Anastasia Kytölä, FinnGen, Mary Pat Reeve)



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?

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