Expanding the OMOP Common Data Model in Accord with Federal Rules for Hospital Price Transparency and Transparency in Coverage

Dr. Jaan Altosaar Li
One Fact Foundation
University of Tartu University
jaan@onefact.org

Michele Tadiello
One Fact Foundation
michele@onefact.org

Jacob S. Zelko
Northeastern University, Roux Institute
One Fact Foundation
jacob@onefact.org

ABSTRACT. Within the OHDSI and general health informatics community, the OMOP CDM serves as an indispensable asset in providing a standardized framework to harmonize Real World Data (RWD) into a consistent structure. As the OMOP CDM continues to gain widespread adoption, researchers are empowered to perform analyses that can generate insights to assist policy decisions. By enforcing a common data representation, the OMOP CDM lends itself strongly to collaborative endeavors and supports deeper investigations into a variety of research domains. While the OMOP CDM supports multiple domains such as diagnoses, drug exposures, and demographic information, its RWD coverage is not complete. In particular, while the OMOP CDM does have some support for analyzing costs associated with patient care through the "Health Economics Data Tables", it is not sufficiently defined to examine prices on a per provider basis, This presents a gap for researchers or policy analysts who desire to assess healthcare economic impacts, cost variations, and value-based decision-making. To address this gap, we present preliminary efforts to develop support for this data species within the Payless Health Common Data Model (CDM), an extension of the OMOP CDM resulting from presentations with the open source working group and the common data model working group related to the limitations of nonprofit organizations in working with open source software and the solutions we have built at One Fact Foundation. The is designed specifically to accommodate detailed cost-related data by incorporating mappings and additional fields. By integrating further cost-related information, the PH-CDM facilitates in-depth cost variations analysis, reimbursement policy examination, and investigation of cost influences on healthcare outcomes.

METHODS

To augment the OMOP CDM, we collected price sheets from 4,025 out of 5,137 hospitals in the United States to prototype this expansion process. Leveraging the existing OMOP CDM "Cost" table as a foundation, we mapped new columns and concepts to accommodate the pricing, reimbursement, and payment details. Additionally, we interviewed several representatives of various health systems to inform our process of developing new fields. An example of such mappings are presented in Figure 1.

Name	Description
`cash_price`	Cash price in accordance with a Federal Rule.
`minimum_reported_price`	Minimum reported price in accordance with a Federal Rule.
`maximum_reported_price`	Maximum reported price in accordance with a Federal Rule.

FIGURE 1. Example additions to the OMOP Common Data Model in accord with the Federal Rule, which requires hospitals to report their prices using data dictionaries provided by the Centers for Medicaid & Medicare Services (CMS 2023a, CMS 2023b).

Standardizing and visualizing clinical data repositories that include price transparency information in compliance to the federal rule. We use the data build tool ("Dbt", n.d.) to standardize clinical data and version control the resulting static, compressed files that are compatible with Amazon Simple Storage Service (S3), which is compliant with the Health Insurance Portability and Accountability Act. Static files in standardized format on Amazon S3 benefit from scalable map-reduce computations that can be parallelized over S3 workers using duckdb (Raasveldt and Mühleisen 2019), an embedded analytical database engine that scales structured query language (SQL) queries across terabytes of data.

Label Studio is an open source data labeling platform that we have scaled to train natural language processing models such as ClinicalBert to analyze over five million records with human clinicians annotating text data at a large academic medical center (Tkachenko et al. 2022). Similarly, we have also used this to implement a labeling workflow for radiologists and demonstrated expert-level imaging diagnosis performance of an open source computer vision model for detecting pediatric upper extremity fractures ("Childfx (Soft Launch; Submission under Review)" 2023). It is important to include financial information in accord with the Federal Rule in such models prior to deployment to assess the likelihood of confounding that is already reported (Zech et al. 2018) — e.g. this study was able to predict the hospital at which an image was taken, highlighting the necessity of including Federal Rule price transparency information to assess the level of confounding in observational bioinformatics research.

(Heer and Moritz 2023) will be used to visualize all of the PH-CDM data in a quality assurance pipeline, which will enable health economics outcomes research and the development of artificial intelligence solutions for health care that incorporate reliable information due to the Federal Rule at national scale.

THE PHENOTYPE WORKFLOW

We have previously standardized and visualized of clinical data repositories of the scale of small countries with millions of patient records. (For example, we have worked with a team in Estonia that has trained ClinicalBert (Huang, Altosaar, and Ranganath 2020) on the entirety of the electronic health record of this country to predict hospital readmission.) However, for disorder and disease classification systems or taxonomies that regularly evolve in response to research (such as coronavirus disease and its sequelae), billing revenue cycle management changes, clinical document improvement practices at academic medical centers, tertiary hospitals, and primary and secondary outpatient clinics, it is necessary to regularly revise the written and computational definitions of the observable characteristics of disease—the phenotype—and assess inter-rater reliability of such definitions (which now includes price transparency information). This enables a practitioner of bioinformatics and observational health research to visualize patterns of health and disease to best

improve clinical guidelines, training, and deployed artificial intelligence solutions free of unobserved confounding due to insurance status or inability to pay. While the OMOP CDM is excellent for this purpose, after several discussions with the phenotype working group, open source working group, and common data model working groups, it became clear that the Payless Health Common Data Model (PH-CDM) would be necessary to expand the OMOP CDM to include prices and other information that is necessary for health economics outcomes research at scale.

For this, we have built the phenotype workflow (Zelko et al. 2023) in collaboration with the phenotype working group in the OHDSI community. The phenotype workflow involves creating a reproducible and replicable open source pipeline for retrospective observational studies, which is a prerequisite for deploying artificial intelligence such as ClinicalBERT into the point of care.

The phenotype workflow requires subject matter experts such as bioinformaticions and clinicians to work alongside research scientists to devise definitions of health, disorder, and disease, and to then annotate patient records in accord with these definitions. Next, the definitions are revised to achive high inter-rater reliability. This has enabled us to validate this workflow by conducting network analyses of tens of millions of patient records using the claims databases at our data partners that include OHDSI.

FIGURE 2. The phenotype workflow can help practice data thinking to best validate definitions of health and disease that are used to train artificial intelligence models, and depend on the Payless Health Common Data Model presented here to enable assessment of phenotype definitions due to typically-unobserved confounders such as inability to pay or dependence on cash price for low-socioeconomic status patients (Zelko et al. 2023).

RESULTS

With these new fields added into the PH-CDM, Payless Health is well-equipped to continue development of the general OMOP CDM. This work can inform discussions on health transparency and care quality within OHDSI working groups (such as the OMOP CDM Workgroup or Health Equity Workgroup) on how researchers within OHDSI can utilize their own data sources alongside the novel expansions produced through development of the PH-CDM. Finally, all code that was developed to create and support these new mappings are open sourced to promote further adoption from other research groups using the phenotype workflow (Zelko et al. 2023) and the Payless Health Common Data Model.

CONCLUSION

The expansion of the OMOP CDM to accommodate Hospital Price Transparency and Coverage data represents a significant advancement in leveraging real-world evidence for healthcare decision-making. By addressing the limitations of the existing OMOP CDM, the Payless Health CDM offers researchers an enhanced framework to explore and analyze cost-related information in conjunction with other clinical and population health data. Through the continued development and potential integration of the Payless Health CDM into the formal OMOP CDM, we strive to contribute to the standardization efforts of OHDSI, ultimately enabling the community to conduct more comprehensive and impactful research in the field of healthcare analytics and provide real world insights to assist in policy decision-making.

REFERENCES

CMS. (2023a). *Resources | CMS*. (Retrieved June 15, 2023, from https://www.cms.gov/hospital-price-transparency/resources)

CMS. (2023b, May 30). *Transparency in Coverage*. Centers for Medicare & Medicaid Services. https://github.com/CMSgov/price-transparency-guide

Heer, J., & Moritz, D. (2023). Mosaic: An Architecture for Scalable & Interoperable Data Views. *Private communication and early access provided to one fact foundation - do not re-cite*.

Huang, K., Altosaar, J., & Ranganath, R. (2020). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *Acm conference on health, inference, and learning*. http://arxiv.org/abs/1904.05342

Raasveldt, M., & Mühleisen, H. (2019, June 25). *DuckDB: an Embeddable Analytical Database* [Paper presentation]. In *Proceedings of the 2019 International Conference on Management of Data*. ACM. https://doi.org/10.1145/3299869.3320212

Tkachenko, M., Malyuk, M., Holmanyuk, A., & Liubimov, N. (2022). *Label Studio: Data labeling software*. https://github.com/heartexlabs/label-studio

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018, November 6). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *Plos medicine*, 15(11). https://doi.org/10.1371/journal.pmed.1002683

Zelko, J. S., Gasman, S., Freeman, S. R., Lee, D. Y., Altosaar, J., Shoaibi, A., & Rao, G. (2023, March 30). *Developing a Robust Computable Phenotype Definition Workflow to Describe Health and Disease in Observational Health Research*. https://doi.org/10.48550/arXiv.2304.06504

ChildFx (Soft launch; submission under review). (2023). (Retrieved May 18, 2023, from https://childfx.com/)

Dbt. (n. d.). Transform data in your warehouse. (Retrieved May 18, 2023, from https://www.getdbt.com/)