Expanding the existing Cadence event sequence visual analysis tool to support the standardized data model OMOP CDM

Natthawut Adulyanukosol and David Gotz

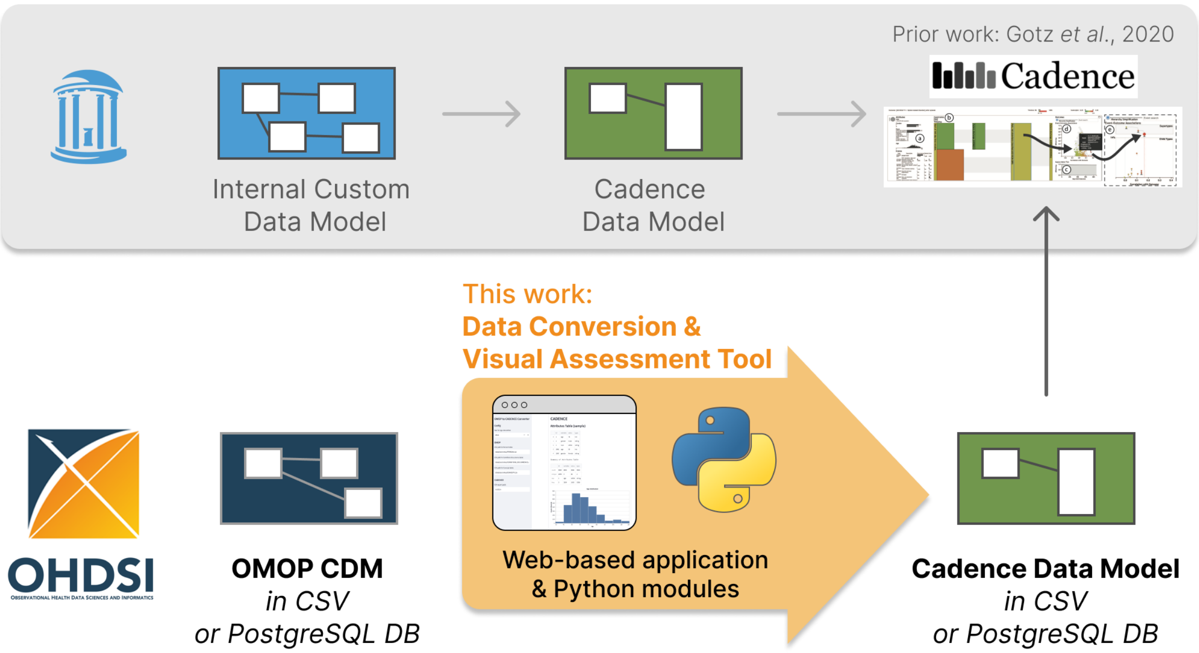


Fig. 1. Data conversion process from data in the OMOP CDM format to be compatible with *Cadence* tool [3]

Abstract—The differences in health data models obstruct the use of analytics tools on new datasets or at other institutions. This work presents a data conversion tool that converts data from the standardized Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to be compatible with the existing *Cadence* event sequence visual analysis tool. The tool is available as a web-based application with an interactive data quality report, which makes the data conversion process more observable visually. Users may detect anomalies and unexpected data distributions in the converted data, and solve the issues prior to uploading the data to *Cadence*. This conversion and visual assessment concept could be applied to other existing analytics tools in healthcare by leveraging the OMOP CDM, and could improve data quality for subsequent analyses.

Index Terms—Data harmonization, Data quality, Visual data mining of EMRs, Longitudinal clinical data

1 Introduction

The differences in health data models from various EMR systems cause significant difficulties in applying visual analytics tools across datasets and institutions. Recent development in the medical data harmonization has introduced several common data models that medical data collected in EMR systems can be converted into. The Observational Health Data Sciences and Informatics (OHDSI) network develops and maintains Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). [5] The OMOP CDM has gained popularity in observational research, and is adopted in many institutions and projects globally, including the US National COVID Cohort Collaborative (N3C) and the US National Pediatric Learning Health System (PDESnet). [1, 4] Medical data that are transformed into the OMOP CDM can be used for further analysis by tools that support the OMOP CDM, for example, HADES PatientLevelPrediction [10], PatientExploreR [2] and Trajectories [6].

Cadence is a web-based temporal event sequence analytical suite with dynamic hierarchical aggregation, which can be used to study event sequence-outcome associations in the medical domain. [3] However, prior to this work, the usage of *Cadence* is limited to a few groups of researchers and *Cadence* not compatible with the OMOP CDM. The users of *Cadence* have to manually convert their data into the format that is required by *Cadence*.

The data transformation and conversion processes have to conform to the specifications of the data models. Given the complexity of medical data models, the development of these processes can be time-consuming. In addition, ad hoc manual data transformation procedure may introduce errors, which can negatively impact the quality of subsequent analyses. To avoid these issues, a data assessment tool can make the conversion process observable, and support early error detection.

In this work, we use *Cadence* as a data conversion target. OMOP

- *Natthawut Adulyanukosol, Carolina Health Informatics Program, University of North Carolina at Chapel Hill, NC, USA. E-mail: na399@unc.edu.*
- *David Gotz, School of Information and Library Science, University of North Carolina at Chapel Hill, NC, USA. E-mail: gotz@unc.edu.*

CDM data is converted into *Cadence* event sequence data format. The data conversion tool is designed to be used on static files and database connections, and providing a data quality report. This conversion concept and its data quality report could be applied to other existing analyics tools by leveraging the OMOP CDM and other standardized data models.

2 Methods

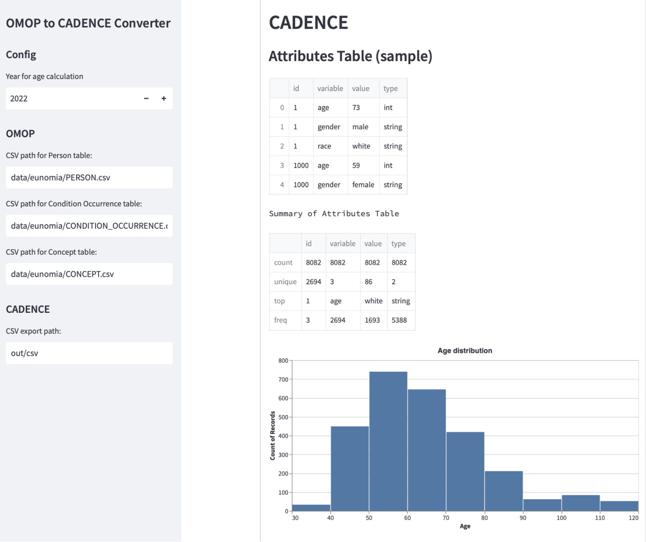
We developed the data conversion and visual assessment tool with Python v3.9. The tool is available as a self-hosted web-based application that users can configure the data conversion settings. The data conversion modules are also executable via command line interface (CLI). The repository of the tool along with its usage manual is publicly available on GitHub at: https://github.com/VACLab/omop-to-cadence-converter.

2.1 Data conversion

The tool imports data tables in the OMOP CDM format as either CSV files via DuckDB [9] or PostgreSQL database connections via psycopg2 [12]. The imported data are converted into the specified *Cadence* format with SQL via DuckDB in-memory engine for CSV files or via psycopg2 on PostgreSQL host server. The converted data are exported as either CSV files or tables on a PostgreSQL database.

2.2 Data quality report

The data quality report is presented visually to the users. We use Streamlit [11] to create the web-based application, and Altair [13] to render the visualization, as displayed in Fig. 2.



OMOP to CADENCE Converter

Config

Year for age calculation: 2022

OMOP

CSV path for Person table: data/eunomia/PERSON.csv

CSV path for Condition Occurrence table: data/eunomia/CONDITION_OCCURRENCE.c

CSV path for Concept table: data/eunomia/CONCEPT.csv

CADENCE

CSV export path: out/csv

CADENCE

Attributes Table (sample)

	id	variable	value	type
0	1	age	73	int
1	1	gender	male	string
2	1	race	white	string
3	1000	age	59	int
4	1000	gender	female	string

Summary of Attributes Table

	id	variable	value	type
count	8082	8082	8082	8082
unique	2694	3	86	2
top	1	age	white	string
freq	3	2694	1693	5388

Fig. 2. A partial screenshot of the web-based application with a data quality report on the right panel. Users can edit configuration settings on the left panel of the application.

3 Results

3.1 Data conversion

We successfully converted sample datasets in the OMOP CDM format to *Cadence* format via both command line interface and the application with graphical user interface. The tested datasets are Eunomia 2.6k as CSV files [7], Synthea 1k on a PostgreSQL database [15], and Synthea COVID 10k on a PostgreSQL database [14]. The converted data can be used for further analyses on *Cadence*.

3.2 Data quality report

The data quality report is presented visually in three forms.

3.2.1 Samples of the original and converted data

We can explore a sample of the original data in the OMOP CDM format to verify that the data are loaded into the application properly. Once the conversion is completed, users can also explore a sample of the converted data in the Cadence format, and check the completeness of expected data columns and values.

3.2.2 Summary statistics of the converted data

The tool reports a descriptive summary statistics of the data in the Cadence format. Any missing data or null values are reported in this statistics.

3.2.3 Distributions of the converted data

The distributions of ages, genders, time of diagnosis, medical code classes, and medical codes are presented as interactive charts. Users can visually spot anomalies or errors in the converted data. For example, in Fig. 2, the ages range from 30 to 120 with a sizable number of patients over the age of 100 years. This distribution might be unlikely and should be fixed prior to subsequent analyses.

4 Discussion

This data conversion tool facilitates the data preparation process from the OMOP CDM format for temporal event sequence analyses in the existing *Cadence* program. *Cadence* specifically requires two data tables that are optimized for efficiency in temporal event sequence pattern finding at scale, while the OMOP CDM stores data in over 30 tables that enable broad analytical purposes. [8] Given the comprehensive data stored in the OMOP CDM, we are able to select only relevant data to be imported into *Cadence*.

We believe that this approach in data conversion from the OMOP CDM may be applied to other existing analytics tools in healthcare to expand the tool adoption. In addition, the development of new medical analytics tools, especially for research settings, may explore and leverage the existing standardized data models as data input for their tools.

4.1 Limitations

We used only synthetic datasets to develop the current version of this tool. Unlike synthetic datasets, actual datasets may have data issues, including missing data and non-standard medical codes, that may not yet be handled properly by this conversion tool.

4.2 Future work

The data conversion tool can be expanded to support to support more queries, other databases and configuration settings, such as cohort specification by demographic information or conditions. Currently, the tool converts the CSV file using in-memory data storage for fast conversion with limited file size. The future version of the tool may support larger CSV files with persistent data storage.

The data quality report may include additional visualizations and automated checks to add coverage for quality control prior to data analyses. In addition, the report could also be produced via the command line interface.

5 Conclusions

The data conversion and visual data quality assessment tool presented in this work expands the compatibility of the existing *Cadence* event sequence visual analysis tool to support data from the OMOP CDM format. The tool is available as a web-based application and Python modules executable via command line interface. This data conversion approach along with the data quality report may be applied to other existing tools to promote the tool adoption by leveraging the OMOP CDM standard, and to improve the quality of the data used for analysis.

References

[1] C. B. Forrest, P. A. Margolis, L. C. Bailey, K. Marsolo, M. A. Del Beccaro, J. A. Finkelstein, D. E. Milov, V. J. Vieland, B. A. Wolf, F. B. Yu, and M. G. Kahn. PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc*, 21(4):602–6, July 2014. Edition: 20140512. doi: 10.1136/amiajnl-2014-002743

[2] B. S. Glicksberg, B. Oskotsky, P. M. Thangaraj, N. Giangreco, M. A. Badgeley, K. W. Johnson, D. Datta, V. A. Rudrapatna, N. Rappoport, M. M. Shervey, R. Miotto, T. C. Goldstein, E. Rutenberg, R. Frazier, N. Lee, S. Israni, R. Larsen, B. Percha, L. Li, J. T. Dudley, N. P. Tatonetti, and A. J. Butte. PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model. *Bioinformatics*, 35(21):4515–4518, Nov. 2019. doi: 10.1093/bioinformatics/btz409

[3] D. Gotz, J. Zhang, W. Wang, J. Shrestha, and D. Borland. Visual Analysis of High-Dimensional Event Sequence Data via Dynamic Hierarchical Aggregation. *IEEE Trans Visual Comput Graphics*, 26(1):440–450, Jan. 2020. Edition: 20190820. doi: 10.1109/TVCG.2019.2934661

[4] M. A. Haendel, C. G. Chute, T. D. Bennett, D. A. Eichmann, J. Guinney, W. A. Kibbe, P. R. O. Payne, E. R. Pfaff, P. N. Robinson, J. H. Saltz, H. Spratt, C. Suver, J. Wilbanks, A. B. Wilcox, A. E. Williams, C. Wu, C. Blacketer, R. L. Bradford, J. J. Cimino, M. Clark, E. W. Colmenares, P. A. Francis, D. Gabriel, A. Graves, R. Hemadri, S. S. Hong, G. Hripscak, D. Jiao, J. G. Klann, K. Kostka, A. M. Lee, H. P. Lehmann, L. Lingrey, R. T. Miller, M. Morris, S. N. Murphy, K. Natarajan, M. B. Palchuk, U. Sheikh, H. Solbrig, S. Visweswaran, A. Walden, K. M. Walters, G. M. Weber, X. T. Zhang, R. L. Zhu, B. Amor, A. T. Girvin, A. Manna, N. Qureshi, M. G. Kurilla, S. G. Michael, L. M. Portilla, J. L. Rutter, C. P. Austin, K. R. Gersing, and the N3C Consortium. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *Journal of the American Medical Informatics Association*, 28(3):427–443, Mar. 2021. doi: 10.1093/jamia/ocaa196

[5] G. Hripcsak, J. D. Duke, N. H. Shah, C. G. Reich, V. Huser, M. J. Schuemie, M. A. Suchard, R. W. Park, I. C. Wong, P. R. Rijnbeek, J. van der Lei, N. Pratt, G. N. Noren, Y. C. Li, P. E. Stang, D. Madigan, and P. B. Ryan. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform*, 216:574–8, 2015.

[6] K. Knnapuu, S. Ioannou, K. Ligi, R. Kolde, S. Laur, J. Vilo, P. R. Rijnbeek, and S. Reisberg. Trajectories: a framework for detecting temporal clinical event sequences from health data standardized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model. *JAMIA Open*, 5(1):ooac021, Apr. 2022. doi: 10.1093/jamiaopen/ooac021

[7] OHDSI. Eunomia, 2022.

[8] OHDSI. OMOP Common Data Model, 2022.

[9] M. Raasveldt and H. Mhleisen. DuckDB: an Embeddable Analytical Database. In *Proceedings of the 2019 International Conference on Management of Data*, pp. 1981–1984. ACM, Amsterdam Netherlands, June 2019. doi: 10.1145/3299869.3320212

[10] J. M. Reps, M. J. Schuemie, M. A. Suchard, P. B. Ryan, and P. R. Rijnbeek. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association*, 25(8):969–975, Aug. 2018. doi: 10.1093/jamia/ocy032

[11] Streamlit Inc. Streamlit The fastest way to build and share data apps, 2022.

[12] The Psycopg Team. psycopg2 - Python-PostgreSQL Database Adapter, 2022.

[13] J. VanderPlas, B. E. Granger, J. Heer, D. Moritz, K. Wongsuphasawat, A. Satyanarayan, E. Lees, I. Timofeev, B. Welsh, and S. Sievert. Altair: Interactive Statistical Visualizations for Python. *Journal of Open Source Software*, 3(32):1057, Dec. 2018. doi: 10.21105/joss.01057

[14] J. Walonoski, S. Klaus, E. Granger, D. Hall, A. Gregorowicz, G. Neyarapally, A. Watson, and J. Eastman. Synthea Novel coronavirus (COVID-19) model and synthetic data set. *Intelligence-Based Medicine*, 1-2:100007, Nov. 2020. doi: 10.1016/j.ibmed.2020.100007

[15] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, and S. McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, Mar. 2018. doi: 10.1093/jamia/ocx079