Powering a Personal Health Record Analytic Environment using the OHDSI CDM and Google Colab

Janos G. Hajagos
Stony Brook University

Background

A personal health record (PHR) should empower a user to better understand and manage their health [1] by combining multiple sources of information. The OHDSI CDM (Common Data Model) is an analytic data model for combining multiple sourced information by aligning elements to a standardized vocabulary. CDA XMLs are a baseline information exchange format supported by electronic health record vendors in the United States [2]. CCDAs, which is a CDA dialect, can be manually downloaded from patient portals and a single CDA document can be exported with the Apple's iOS Health App. This work demonstrates that multiple sourced CDAs from a single person can be converted to the OHDSI schema and analyzed in the interactive Google Colab Python environment.

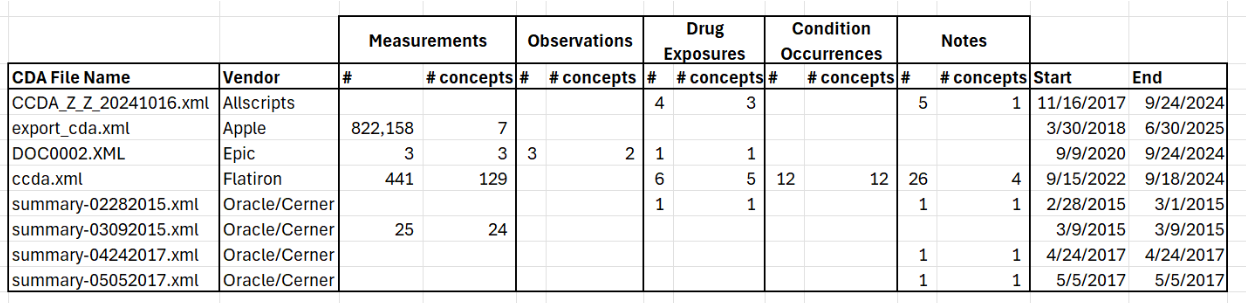
Methods

Multiple XML CDAs from EHR patient portals and Apple's iOS Health App were extracted and uploaded to a secure cloud file share. The data was then mapped to the intermediary (Prepared Source Format) PSF in the Apache Parquet Format and then transformed to the OHDSI CDM version 5.4 with the vocabulary release v5.0 27-FEB-25. The whole pipeline is executed in a Google Colab Python Notebook and does not involve manual software dependency installation. The instructions for running the pipeline are on GitHub [3].

Currently supported extractions from CDA sections include vital signs, laboratory measurements, measurements from devices that connect to Apple's iOS Health App, diagnoses, medications, and clinical notes. Most CDA documents contain coded data in either SNOMED, LOINC, and RxNorm which is supported in the OHDSI vocabulary.

Results

A total of 8 CDA XMLs files from 5 different vendors were converted to the OHDSI CDM covering over 10 years of personal health data. A summary of the conversion is included in Figure 1 and a sample data analysis which combines multiple sources of information is shown in Figure 2.



		Measurements		Observations		Drug Exposures		Condition Occurrences		Notes			
CDA File Name	**Vendor**	**#**	**# concepts**	**#**	**# concepts**	**#**	**# concepts**	**#**	**# concepts**	**#**	**# concepts**	**Start**	**End**
CCDA_Z_Z_20241016.xml	Allscripts					4	3			5	1	11/16/2017	9/24/2024
export_cda.xml	Apple	822,158	7									3/30/2018	6/30/2025
DOC0002.XML	Epic	3	3	3	2	1	1					9/9/2020	9/24/2024
ccda.xml	Flatiron	441	129			6	5	12	12	26	4	9/15/2022	9/18/2024
summary-02282015.xml	Oracle/Cerner					1	1			1	1	2/28/2015	3/1/2015
summary-03092015.xml	Oracle/Cerner	25	24									3/9/2015	3/9/2015
summary-04242017.xml	Oracle/Cerner									1	1	4/24/2017	4/24/2017
summary-05052017.xml	Oracle/Cerner									1	1	5/5/2017	5/5/2017

Figure 1. A table of record counts from each CDA file converted to the OHDSI schema for the supported data domains (n = 1). For each data domain the unique number of concepts is determined, and a date range is calculated.

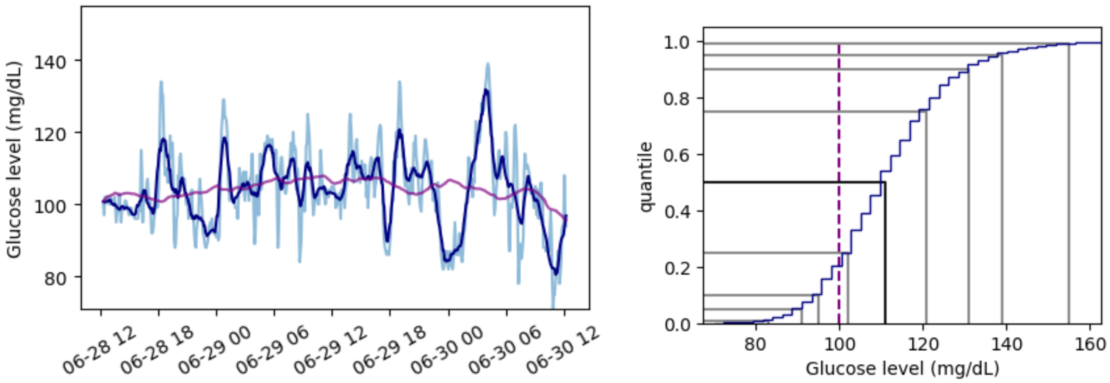


Figure 2. The first plot is of continuous glucose measured by an OTC device over the last 48 hours with actual values (light blue), one hour moving average (dark blue), and 12 hour moving average (purple solid line). The second plot shows the cumulative distribution of glucose levels for the past 90 days (excluding the last 48 hours) based on 21,272 separate measurements. The purple dashed line shows the last blood glucose measured in a clinical setting. Other lines show standard quantiles ranging from 0.01 to 0.99. Data was extracted automatically from CDA documents and converted to the OHDSI CDM and analyzed in the Google Colab environment.

Conclusion

This work demonstrates that a single person's health data obtained manually from multiple sources can be converted to the OHDSI CDM and analyzed in a single database. This preliminary conversion covered multiple EHR sources and patient captured data devices. The current OHDSI vocabulary does not include an OHDSI type for data collected by the patient. It is important to distinguish data collected in routine patient care and those collected using personal health devices as they have different sampling frequencies and quality. If the OHDSI CDM will be used for self-collected patient device data it will require significant database scaling due to the large volume of data a single person can generate.

References

1. Berg, L. N. van den et al. The feasibility and usability of a personal health record for patients with multiple sclerosis: a 2-year evaluation study. Front. Hum. Neurosci. 18, 1379780 (2024).
2. D'Amore, J. D. et al. Are Meaningful Use Stage 2 certified EHRs ready for interoperability? Findings from the SMART C-CDA Collaborative. J. Am. Méd. Inform. Assoc. 21, 1060–1068 (2014).
3. https://github.com/jhajagos/PHR2OHDSI