Size comparison of 17 CDM datasets using IRIS tool

Vojtech Huser, MD PhD[1,3], Marc A. Suchard, MD PhD[2,3]
[1]Lister Hill National Center for Biomedical Communication, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[2]{Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, Department of Biostatistics, Fielding School of Public Health} at University of California, Los Angeles, CA, USA
[3]Observational Health Data Sciences and Informatics

Abstract

With availability of multiple Big Data healthcare datasets, researchers may need a way to compare them in terms of size and quality. We present a new tool, called Iris, which allows comparison of datasets that follow the Common Data Model (CDM) currently maintained by the Observational Health Data Sciences and Informatics (OHDSI) collaborative. Iris uses a set of defined count measures to describe the size of a CDM dataset. This poster presents results of applying the Iris tool (version 1) on 17 datasets that represent large claims databases, national Electronic Health Record (EHR) databases, national large-scale research surveys, and repositories of academic medical centers or health information exchanges. Development of Iris is motivated by efforts to characterize a "gold standard patient" or a "minimum data patient" for observational secondary research utilizing EHR data. We expect significant future evolution of the tool and the included data size measures.

Introduction

Transforming healthcare data into a Common Data Model (CDM) allows execution of analyses on several large healthcare datasets. During analysis planning, analyst often need to understand differences in datasets across several data partners. Using the CDM maintained by the Observational Health Data Sciences and Informatics (OHDSI) collaborative,[1] we developed a tool, called Iris, which allows a quick generation and comparison of several well defined data size metrics. In this poster, we present a comparison of 17 datasets that represent large claims databases, national Electronic Health Record (EHR) databases, national large-scale research surveys, and repositories of academic medical centers or health information exchanges.

Methods

The Iris tool was first released in March 2015 as a Structured Query Language (SQL) script and later wrapped into an R package for easier execution. Iris code is open source (available at github.com/OHDSI/iris) and can be executed within R after a single install command – devtools::install_github("OHDSI/Iris"). The core code consists of a set of SQL queries that compute counts, such as 'Count of patients with at least one diagnosis and one procedure'; measure D3). Iris is inspired by our earlier effort to perform dataset size comparisons.[2]

Iris currently provides the following measures: count of events (G1), count of patients (G2), count of patients with at least one diagnosis and one medication (D2), count of patients with at least one diagnosis and one procedure (D3), count of patients with at least one diagnosis, medication and observation (D4), and count of deceased patients (D5).

Development of Iris was motivated by discussions within the OHDSI community about "gold standard patient" or "minimum data patient". Some datasets may contain a large number of patients with data limited to only one domain. For example, a data partner may include in the CDM dataset all data from the regional immunization registry. Such inclusion will greatly increase the number of patients in the dataset, however there will be no diagnoses, procedures, lab results or medications recorded for most immunization registry patients because they receive care outside the integrated delivery network. Similar heterogeneous pattern may occur in a health information exchange (HIE) dataset where widely differing level of patient data may be provided by the local HIE participants. For calculations of disease or event prevalence (% of patients with a given EHR event), inclusion of all possible patients may lead to biased population level prevalence estimates.

The set of measures currently computed by Iris is not meant to be comprehensive but serves as a starting point for more complex dataset comparisons in future versions of Iris or in OHDSI Achilles tool.[3] We used on purpose a strategy of a limited set of measures and an early release of Iris to allow CDM adopters to see their dataset quantified using Iris and provide additional measures that they would want to add to future version of Iris and why they consider them important.

Results

Table 1 shows values for all current Iris measures for the analyzed 17 datasets (with some sites providing Iris data for multiple datasets). To encourage broader site participation for our comparison, we masked the names of the individual datasets (and sites) and refer to datasets using arbitrary IDs. We received informal positive feedback from Iris users that the counts provide an interesting initial insights about their dataset with respect to the "minimum data patient" problem. In terms of request for additional measures, one site requested modifications for measure D4 that uses OBSERVATION table that we plan to address in future Iris releases. Iris version 1 was designed to run on both CDM version 4 and version 5 and for all other measures (except D4) the same logic can be used on both CDM versions. Because version 5 introduced a separate table for a set of laboratory data items (defined in OHDSI Vocabulary using the concept domain mechanism), a new measure D4b can potentially asses patients with at least one diagnosis, medication and lab measurement.

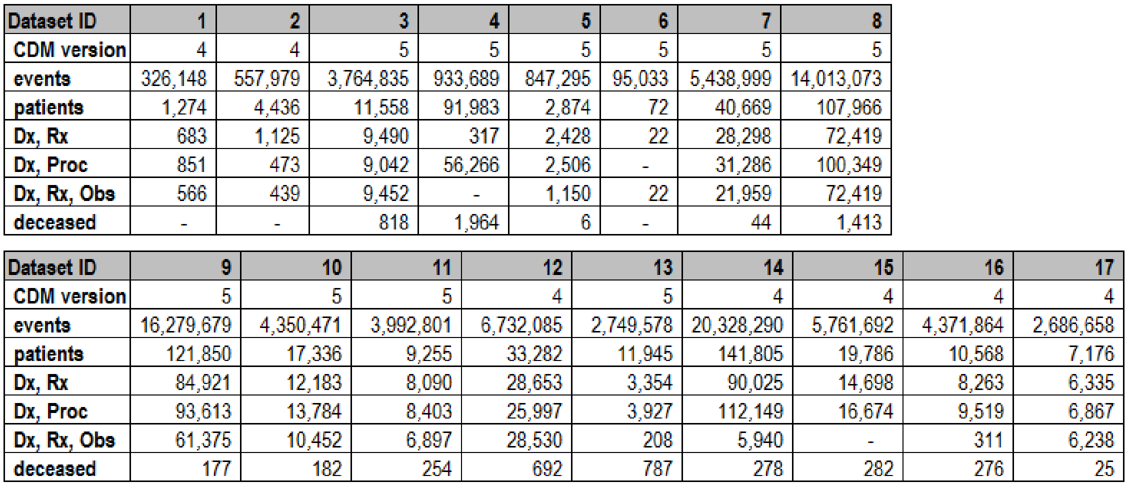
Review of results in Table 1 shows quickly that some datasets contain no data on patient deaths or observations. For researchers conducting research in pharmacovigilance, Iris shows significant differences in counts of G1 patients (all patients) to D2 patients (1+[Dx, Rx] patients). For example, dataset #8 with 107.9 million G1 patients drops to 72 million D2 patients (33% decrease).

Conclusion

We have created a new OHDSI tool that allows elementary data size comparison between CDM datasets, and we generated size comparisons for 17 CDM datasets.

Acknowledgement: We would like to thank Mary Boland, Shang Ning, Patrick Ryan, Anthony Sena, Jon Duke for their contribution to this work. This research was supported in part (VH) by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and Reagan-Udall Foundation for the FDA project IMEDS-SA-0011. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services.

Table 1. 2-part table showing IRIS results for 17 compared datasets (counts indicate thousands; except the version row).



Dataset ID	1	2	3	4	5	6	7	8
CDM version	4	4	5	5	5	5	5	5
events	326,148	557,979	3,764,835	933,689	847,295	95,033	5,438,999	14,013,073
patients	1,274	4,436	11,558	91,983	2,874	72	40,669	107,966
Dx, Rx	683	1,125	9,490	317	2,428	22	28,298	72,419
Dx, Proc	851	473	9,042	56,266	2,506	-	31,286	100,349
Dx, Rx, Obs	566	439	9,452	-	1,150	22	21,959	72,419
deceased	-	-	818	1,964	6	-	44	1,413

Dataset ID	9	10	11	12	13	14	15	16	17
CDM version	5	5	5	4	5	4	4	4	4
events	16,279,679	4,350,471	3,992,801	6,732,085	2,749,578	20,328,290	5,761,692	4,371,864	2,686,658
patients	121,850	17,336	9,255	33,282	11,945	141,805	19,786	10,568	7,176
Dx, Rx	84,921	12,183	8,090	28,653	3,354	90,025	14,698	8,263	6,335
Dx, Proc	93,613	13,784	8,403	25,997	3,927	112,149	16,674	9,519	6,867
Dx, Rx, Obs	61,375	10,452	6,897	28,530	208	5,940	-	311	6,238
deceased	177	182	254	692	787	278	282	276	25

References

1. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in health technology and informatics.* 2015;216:574-578.
2. Huser V, Cimino JJ. IDR Snapshot: Quantitative Assessment Methodology Evaluating Size and Comprehensiveness of an Integrated Data Repository. *AMIA Translational Bioinformatics Summit 2012* 2012.
3. OHDSI. Observational Health Data Sciences and Informatics Wiki: Achilles Documentation. 2015; http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:achilles. Accessed June 22, 2015.