Assessing eMERGE Common Data Elements on the All of Us Data

Xinzhuo Jiang, MS[1], Ning Shang, PhD[1], Anna Ostropolets, MD[1],
Chunhua Weng, PhD[1], Karthik Natarajan, PhD[1]
[1]Department of Biomedical Informatics, Columbia University, New York, NY

Background

The *All of Us* (AoU) Research Program is a nationally launched precision medicine program which aims to create a diverse health database with over 1 million participants. The data set includes demographic information, enrollment registration data, EHR, participant-provided information (PPI) via questionnaires and surveys, physical measurement and genomic information[1]. It is stored in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).

There are no established metrics to evaluate how comprehensive the AoU dataset is for conducting observational research. This study characterized the AoU data set using common data elements (CDEs) extracted from PheKB[2] phenotype definitions in order to assess the AoU data set.

Research Category

Phenotyping and Data Quality Assessment

Methods

Köpcke et al. analyzed the presence of common data elements in electronic health records in order to compare the eligibility criteria defined in clinical trial protocols[3]. Applying this idea, the primary task here was to evaluate the AoU data set using CDEs in phenotype definitions. In PheKB, there are 53 validated phenotypes and each phenotype algorithm is defined by concepts[4]. The semantic collections of similar concepts within a phenotype are called *concept sets*. Each concept set contains a group of standard coding schema concepts[5]. One concept set could occur among multiple phenotypes as shown in Figure 1. We created a list of the concept sets. In addition, we added basic variables found in the PheKB data dictionary such as age, gender and visit to the concept set list in order to better characterize cohorts. The final list of variables and concept sets will be referred to as the list of CDEs. Using the CDEs, we examined which ones frequently occurred in eMERGE phenotypes. To evaluate the capability of AoU data set for running phenotype studies, we calculated the prevalence of each CDE in the data set. We also compared the prevalence of CDEs using both standard coding schema concepts and source coding schema concepts.

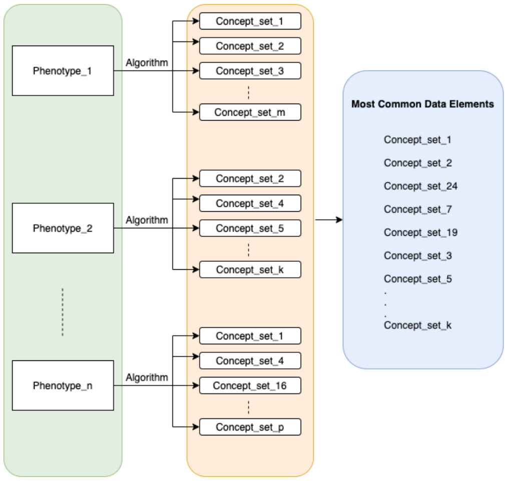


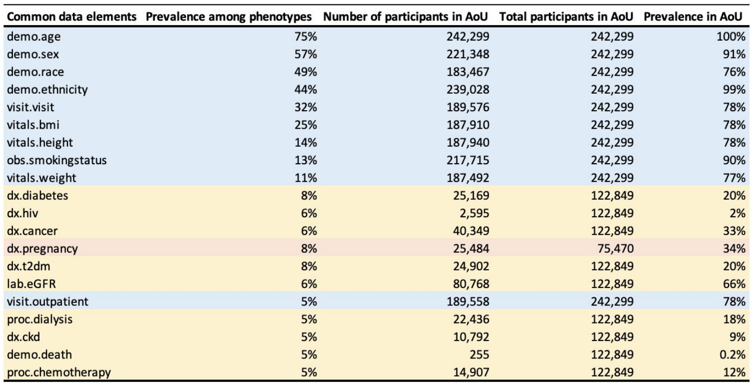
Figure 1. The extraction process of common data elements from eMERGE phenotypes

Results

In AoU data set, there are 242,299 distinct participants (57% female and 35% male). 55% (n=122, 849) of the participants have EHR data.

Among 63 eMERGE phenotypes and 651 data elements, we identified the top 20 most frequent CDEs across phenotypes as listed in Table 1. The prefix stands for the clinical domain of the variable, e.g. 'vitals.bmi' is a vitals measurement, specifically Body Mass Index. The column 'Prevalence among phenotypes' lists the occurrence of each CDE across 63 phenotypes; the column 'Number of participants in AoU' lists the number of distinct participants who had at least one observation of the concept ids of the CDE.

We separated the CDEs into three groups. For the blue group, the ‘Total participants in AoU’ is 242,299 because these CDEs consist of information which could be accessed from every participant’s registration data or EHR data. For the yellow group, the ‘Total participants in AoU’ is 122,849 because these CDEs are only available from participants who provided EHR. Finally, for the pink group, the ‘Total participants in AoU’ is 75,470 because pregnancy prevalence is specific to the female population. The column ‘Prevalence in AoU’ is the ratio of ‘Number of participants in AoU’ to ‘Total participants in AoU’.



Common data elements	Prevalence among phenotypes	Number of participants in AoU	Total participants in AoU	Prevalence in AoU
demo.age	75%	242,299	242,299	100%
demo.sex	57%	221,348	242,299	91%
demo.race	49%	183,467	242,299	76%
demo.ethnicity	44%	239,028	242,299	99%
visit.visit	32%	189,576	242,299	78%
vitals.bmi	25%	187,910	242,299	78%
vitals.height	14%	187,940	242,299	78%
obs.smokingstatus	13%	217,715	242,299	90%
vitals.weight	11%	187,492	242,299	77%
dx.diabetes	8%	25,169	122,849	20%
dx.hiv	6%	2,595	122,849	2%
dx.cancer	6%	40,349	122,849	33%
dx.pregnancy	8%	25,484	75,470	34%
dx.t2dm	8%	24,902	122,849	20%
lab.eGFR	6%	80,768	122,849	66%
visit.outpatient	5%	189,558	242,299	78%
proc.dialysis	5%	22,436	122,849	18%
dx.ckd	5%	10,792	122,849	9%
demo.death	5%	255	122,849	0.2%
proc.chemotherapy	5%	14,907	122,849	12%

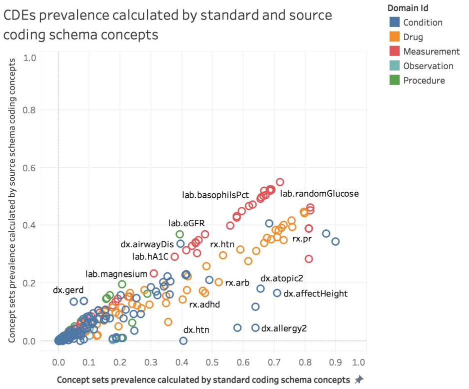


Table 1. Top 20 CDEs and prevalence in *All of Us* data set

Figure 2. Prevalence computed by standard and source concepts

In Table 1, the data coverages of physical measurements including BMI, height and weight are around 78%. Diabetes diagnosis, including Type I and Type II Diabetes Mellitus, has 5 occurrences among all phenotypes and over 20% of participants have had related diagnoses. The coverage of most demographic data elements is above 90%. Participants could skip answering some questions, resulting in lower prevalence of the demographic CDEs. In terms of visits, 78% of the participants had at least one visit at the health provider organization where recruitment occurred and almost everyone in this cohort had at least one outpatient visit.

Figure 2 shows the comparison between the prevalence calculated by standard coding schema concepts and source coding schema concepts. X axis is the prevalence computed by standard coding schema concepts and y axis is the prevalence computed by source coding schema concepts of the same concept set. For example, the CDE lab.randomGlucose coverage in the AoU data set is 71% for standard coding schema concepts and 54% for source coding schema concepts.

Conclusion

This study applied common data elements from the eMERGE phenotypes as an evaluation criteria to check the data availability of the CDEs on the AoU data set in order to assess feasibility of phenotype studies. We extracted the top 20 CDEs and computed the prevalence of each CDE across the AoU data set. We also studied the prevalence difference calculated by standard and source coding schema concepts. For future study, we will expand this work to exam CDE coverage and build up a threshold to evaluate data completeness from a phenotyping perspective.

Acknowledgments

This project is supported by *All of Us* (1U2COD023196, 1UG3OD023183) and eMREGE (U01HG008680).

References

1. Allofus.nih.gov. 2020. National Institutes Of Health (NIH) — All Of Us. [online] Available at: <https://allofus.nih.gov/> [Accessed 24 March 2020].
2. Phekb.org. 2020. What Is The Phenotype Knowledgebase? | Phekb. [online] Available at: <https://phekb.org/> [Accessed 24 March 2020].
3. Köpcke, F., Trinczek, B., Majeed, R., Schreiweis, B., Wenk, J., Leusch, T., Ganslandt, T., Ohmann, C., Bergh, B., Röhrig, R., Dugas, M. and Prokosch, H., 2013. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Medical Informatics and Decision Making, 13(1)
4. 53 phenotype names: https://drive.google.com/open?id=1-mTqclV75EYLHpfsxLwZCaIMEhWfxEhT
5. Informatics, O., 2020. The Book Of OHDSI. [online] Ohdsi.github.io. Available at: <https://ohdsi.github.io/TheBookOfOhdsi/> [Accessed 25 March 2020].