Assessing and Benchmarking Data Quality and Diversity in the *All of Us*

Lina Sulieman, Karthik Natarajan, Kayla Marginean, Robert Carroll, Paul Harris

Background

With the surge of real-world clinical data collection for research, the necessity of assessing and benchmarking Data Quality (DQ) rises to ensure the utility of the data and improve the reproducibility of research.[1] The *All of Us* program is a national initiative that is collecting Electronic Health Records Data (EHR), surveys, physical measurements, and genetics data from participants that reflect the diversity of the United States. The quality of the data can impact the research credibility, which is crucial, especially given the complexity of the *All of Us* dataset and the diversity of researchers' backgrounds. The objective of this study is to assess the quality and diversity of the EHR in the *All of Us*.

Methods

From the *All of Us* controlled tier launched in March 2022, we extracted participants who have EHR. To identify the prevalence of various phenotypes and assess the diversity in phenotypic cohorts, we implemented the OHDSI phenotype library to extract 212 phenotypes using 763 algorithms that use a combination of conditions, drugs, measurements, visits, and procedures. We compared the *All of Us* prevalence of the phenotypes that are considered the leading cause of death in the US to their prevalence as reported by the Center for Disease and Control (CDC).

To assess the quality of EHR data, we ran the OHDSI DQ dashboard package and extracted plausibility, conformance, and completeness metrics. Plausibility measures the extent to which the values agree with internal and external knowledge. Conformance quantifies the percentage of the dataset that complies with standards and constraints (e.g., allowable ranges of lab values). Completeness measures the percentage of data that is expected to be present (e.g., the percentage of rows in the drug table with dose value). Moreover, we added another completeness metric to the one defined by the OHDSI tool which is the existence of core measurements. We calculated the percentage of participants who have core measurements: height, weight, Body Mass Index (BMI), cholesterol, and heart rate. Since the *All of Us* receives data from multiple sites, we extracted the core measurement percentages per site.

Results

The *All of Us* controlled tier included 331,382 participants where 76% were considered underrepresented in biomedical research and 55.81% were white compared to 76.30% white in the US general population based on the census data. Around 60% of participants were female and selected women as their gender identity and 0.12% were transgender. Around 76% of participants had any EHR data, where 45% participants had height and 46% had weight measurements in the EHR.

Of the 224,507 participants with EHR data, the OHDSI phenotype library identified 223,018 participants with at least one of the 212 EHR-based phenotypes. The mean prevalence of the US leading causes of death phenotypes were close to or slightly higher than the prevalence reported by CDC except for Alzheimer's disease, suicide, and influenza as Figure 1 shows. Most of the OHDSI phenotypes had a higher percentage of white participants compared to the percentage of enrolled white participants (Figure 2). However, 413 (55%) of the cohorts had 60% or lower white participants. Around 80% of the *All of Us* dataset values were plausible, 74% conformed with the dataset standards, and 78% of the data entries were complete, as Figure 3 depicts. Both completeness and conformance had values higher than 75% in the verification and validation categories (Figure 3). The most common plausibility issues were out-of-range values (e.g., Calcium in serum/plasma lower than 7) and gender-specific conditions and procedures

concepts that were in participants' records with the opposite gender (e.g., "Deliveries by cesarean" in male participants' record). Using non-standard and non-existing concepts were the most common conformance issues such as drug route concepts do not exist in the defined concept list. Null or zero values were the most common completeness issues. The completeness of core EHR physical measurements in all sites ranged between 3%-100% for height, and weight, and between 11% to 100% for BMI and heart rate.

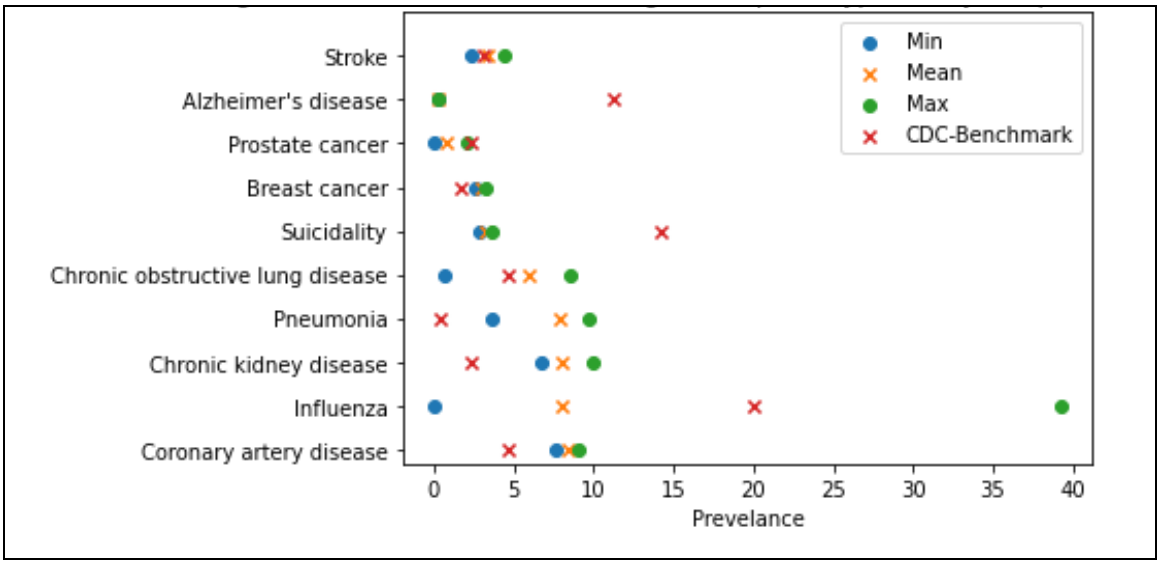

Figure 1. The prevalence of phenotypes that are the leading causes of death as reported by the CDC. Each phenotype had more than one algorithm. We reported the minimum, maximum, and mean prevalence for all algorithms per phenotype

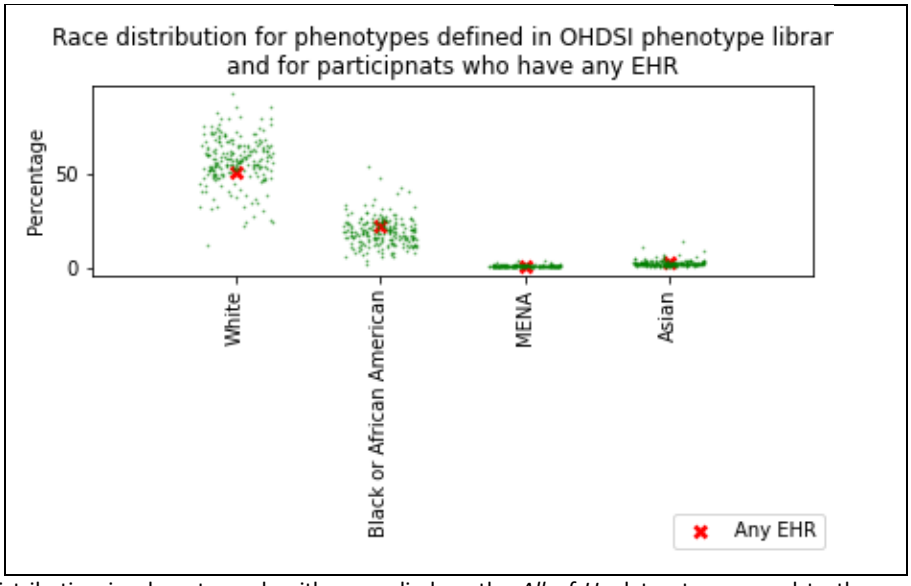

Figure 2. Race distribution in phenotype algorithms applied on the *All of Us* dataset compared to the race distribution for participants with any EHR

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1821	192	2013	90%	8	279	287	3%	1829	471	2300	80%
Conformance	492	189	681	72%	92	12	104	88%	584	201	785	74%
Completeness	299	87	386	77%	13	2	15	87%	312	89	401	78%
Total	2612	468	3080	85%	113	293	406	28%	2725	761	3486	**78%**

Figure 3. The plausibility, conformance, and completeness metrics in the All of Us controlled release

Conclusions

We implemented the OHDSI tools to assess the DQ and diversity in the *All of Us* EHR by replicating OHDSI phenotype algorithms and quantifying three DQ metrics: plausibility, conformance, and completeness. Our analysis demonstrated that the prevalence of some phenotypes in *All of Us* was higher than the prevalence reported by the CDC, except for Alzheimer's, suicide, and influenzas. This could be due to implementing simpler versions of phenotype algorithms. Additionally, recruitment at each site might influence the disease prevalence within *All of Us* (i.e., multiple sites might recruit participants from their breast cancer clinic). The observed lower prevalence of Alzheimer's Disease is expected since inclusion criteria for the *All of Us* program currently require a decisional capacity to consent. The racial distribution in the *All of Us* demonstrated more diversity compared to other research repositories.

The EHR quality seemed within an acceptable range. The conformance of the dataset might be low due to mapping differences in EHR sites. The very low plausibility in validation might occur due to the external source used in the evaluation which requires more investigation.

Data quality and diversity are essential factors that can improve clinical research reproducibility. The *All of Us* is a national initiative that is collecting health data from diverse participants for research and is available for diverse researchers. Our analysis demonstrated the diversity of the *All of Us* EHR and highlights quality issues that need to be investigated.

References/Citations

1. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. Vol. 27, Journal of the American Medical Informatics Association. 2020.