Using Cohort Diagnostics to Assess the Phenotypic Data Quality in All of Us Research Program

Lina Sulieman[1], Karthik Natarajan[2]

[1]Vanderbilt University Medical Center, Nashville, Tennessee, [2]Columbia University Irving Medical , New York City, New York

Background

The quality of clinical data used in research can influence the utility of the data and the reproducibility of clinical research.[1] All the current approaches of assessing data quality are not phenotype specific. Providing methods and techniques to assess the data quality based on the phenotype for which the data will be used can ensure the reproducibility of the research. OHDSI cohort diagnostics is one the few tools that conduct phenotypic-specific diagnostics. Comparing the results of applying Cohort Diagnostics on a research repository and comparing those results to other datasets can ensure the fitness-of-use of the dataset to conduct research for a given phenotype. The *All of Us* Research Program is a national initiative that is collecting Electronic Health Records Data (EHR), surveys, and genetics data from historically underrepresented population in biomedical research. The quality of the data in the *All of Us* Research Program can impact research credibility. The objective of this study is to utilize OHDSI's Cohort Diagnostics to assess the phenotypic data quality in the All of Us Research Program, focusing on breast cancer.

Methods

We applied cohort diagnostics on the *All of Us* Research Program controlled tier released in March 2022 for breast cancer phenotype included in OHDSI phenotype library9. Using the output, we identified the percentages of overlaps between cohorts calculated by the number of subjects in both cohorts over the number of subjects in either cohort. We extracted the incidents rates, time distributions, and covariant. We compared breast cancer cohort analyses between the All of Us dataset and multiple datasets published on https://data.ohdsi.org/CohortDiagnosticsBreastCancer/.

Results

The *All of Us* controlled tier included 331,382 participants, where 55.81% were white and 60% of participants were female. Applying breast cancer cohorts from the phenotype library, two algorithms with identifiers Cohort-1: 4112853001 and Cohort-2: 4112853002 identified X and Y respectively. The number of participants included in both cohorts was 5905 which is 79.69% of participants included in either cohort. All participants identified in Cohort-1 were included in Cohort-2. Both cohorts included female participants as well as participants identified with different gender other than female. No cases were reported in participants who were younger than 30 years old, as Figure 1 shows. The trend in incident rates across age groups matches the trend observed in the ten other datasets on which breast cancer diagnostics was applied.

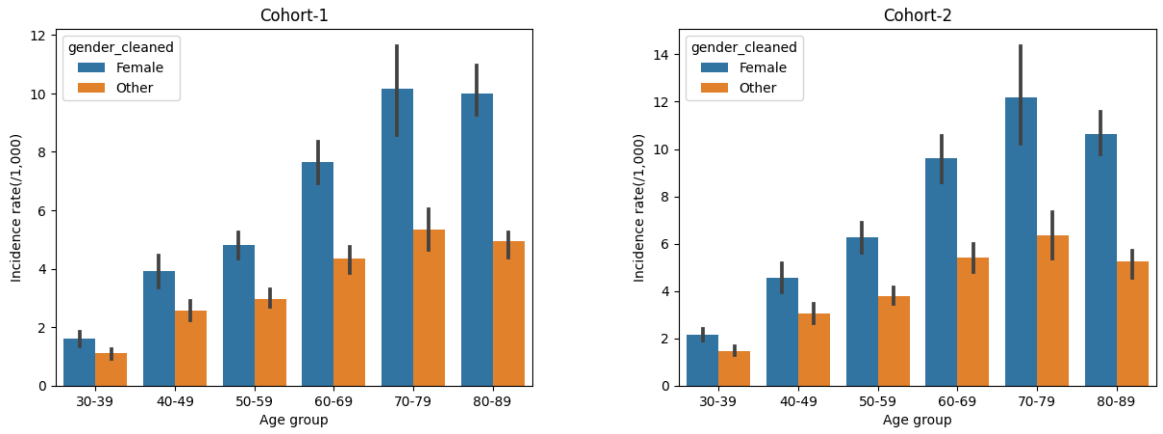

Figure 1. Incident rates of breast cancer for participants identified by breast cancer Cohort-1 and Cohort-2

The median of time in days before and after the index diagnosis of breast cancer in the All of Us is higher than the other datasets. The biggest difference in median time can be seen in the times prior to index time. The closest dataset in time distributions was Clinical Practice Research Datalink (CPRD) followed by Columbia University Irving Medical Center (CUIMC).

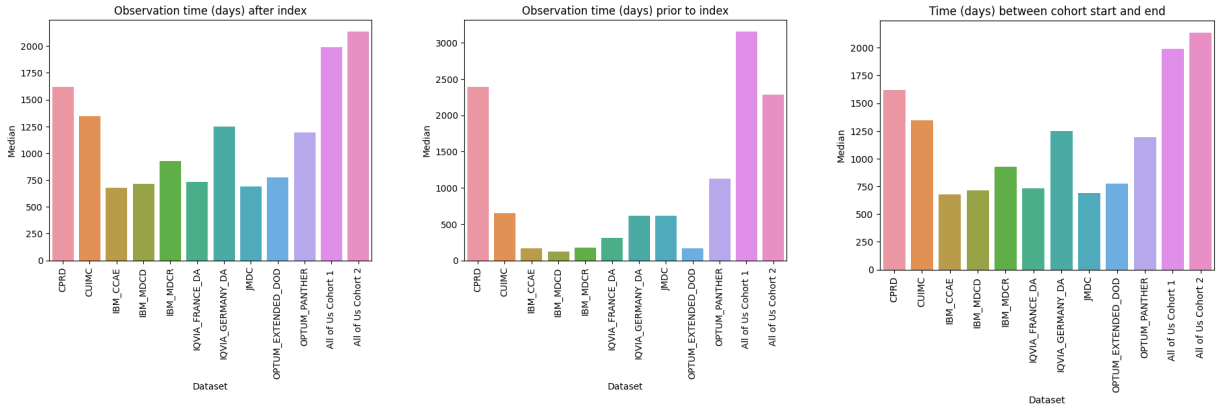

Figure 2. Comparing the median time distribution prior, before, and between breast cancer cohort time between the *All of Us* cohort and OHDSI datasets

We compared the temporal characteristics for weight, height, tamoxifen, letrozole, Primary malignant neoplasm of female breast code, Hemoglobin, and Neutrophils/100 leukocytes covariates. For weight and height, the mean increased by at least 0.15 in "Start -365 to -31" and "Starts 31 to 365" periods, as Table 1 shows. There was no breast cancer condition reported prior to index date. The mean value of Tamoxifen and Letrozole, which are drugs used to treat breast cancer, increased by 10 folds in "Starts 31 to 365" periods. The mean values of blood tests such as Hemoglobin and Neutrophils had doubled in "Starts 31 to 365" period as Table 1 depicts. Table 1 lists the Temporal Characterization of mean values for selected covariates.

Table 1. Temporal Characterization of mean values for selected covariates that are part of breast cancer cohort building and clinical management

	Start -365 to -31		Start -30 to -1		Start 1 to 30		Starts 31 to 365	
	Cohort-1	Cohort-2	Cohort-1	Cohort-2	Cohort-1	Cohort-2	Cohort-1	Cohort-2
Primary malignant neoplasm of female breast	0	0	0	0	0.4276	0.4947	0.5907	0.5955
Height	0.4115	0.3497	0.2041	0.1798	0.3538	0.3368	0.6032	0.5781
Weight	0.2909	0.2526	0.1561	0.1408	0.2476	0.2457	0.4176	0.4190
Tamoxifen	0.0119	0.0105	0.0042	0.0040	0.0085	0.0092	0.0936	0.0911
Letrozole	0.0020	0.0016	0.0017	0.0019	0.0076	0.0076	0.0566	0.0553
Hemoglobin	0.3634	0.3066	0.1272	0.1157	0.2793	0.2661	0.5687	0.5440
Neutrophils/100 leukocytes	0.0786	0.0655	0.0285	0.0247	0.0650	0.0592	0.1573	0.1412

Discussion

We leveraged the Cohort diagnostic to assess the quality of breast cancer cohort extracted from the *All of Us* research Program. Our analysis showed an expected trend in breast cancer in different age groups. There is a higher-than-expected incident rates in participants identified as Other than female. This might be due to having more than one way or category to assign sex or gender identity in the *All of Us* dataset. This analysis demonstrate that researchers should apply a careful extraction and classification of participants sex when extracting any cohort. The trend in temporal characteristics for covariates matches what we expect to see in clinical setting after breast cancer diagnosis. For example, the mean of drug used to treat breast cancer was much higher than the mean before diagnosis. Measurements that are used to evaluate patient health and prepare chemotherapy medications such as height, weight, and blood tests increased as well after the diagnosis. This increase is expected since those measurements are taken periodically after diagnosis as part of treatment.

Conclusion

Assessing the quality of cohort that will be used in biomedical research is essential. As the All of Us data grows and more researchers access and use this source, we need to examine the quality of data on the phenotype level. General data quality metrics will miss important factors that are essential for the credibility of phenotype research. Our feasibility analysis demonstrated that Cohort Diagnostic tool can be repurposed to assess the quality of phenotype cohorts.

Reference

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.