Characterizing Phenotype Descriptions in *All of Us* Publications

Emily Clark, MPH[1], Matthew Spotnitz, MD, MPH[1], Lew Berman, PhD, MS[1], John Giannini, PhD[1], Yechiam Ostchega, PhD, RN[1], Lakshmi Priya Anandan, MPH[1]

[1]*All of Us* Research Program, Office of the Director, National Institutes of Health, Bethesda, Maryland

Background

Phenotypes are an essential component of observational healthcare research, and the basis for a myriad of patient cohorts. However, there is substantial variability in the methods used for defining and describing phenotypes. Clearly defined phenotypes are important in observational research to allow researchers to build on prior work without introducing error or bias in phenotypic variation. Having a common language for discussing phenotypes, which can be regulated in journal requirements, may improve the clarity, reproducibility, and rigor of phenotyping. The *All of Us* Research Program provides researchers with a rich platform to define phenotypes and align specific use cases with the OMOP CDM standard for broad use and adoption.[1] The *All of Us* Researcher Workbench (RW) provides a valuable opportunity to test and apply data standards by researchers of varying locations, expertise, and clinical focus areas.

Methods

The study design of this paper was adopted from Brandt et al. who developed a framework for describing phenotypic variability across datasets.[2] Our aim was to evaluate the variability in phenotypic definitions in *All of Us* publications. The program tracks publications on Pubmed that mention "The *All of Us* Research Program" and lists them on the *All of Us* Publications page, which we used for our review.[3] We manually reviewed the list of published papers from the program's inception until the end of December 2022. Papers were included if authors studied a phenotype, and papers with multiple phenotypes were evaluated on the main phenotype. We excluded papers that described program operations or a genome-wide association study (GWAS) only. We described which source codes and data domains were used in the definition of a study's main phenotype.

Results

A total of 130 papers were published between the program's inception and the end of 2022. Each of these papers were reviewed manually, and of those, 69 (53%) were included in this analysis. Of those included, there were 44 (64%) phenotypes that mentioned International Classification of Diseases (ICD) diagnostic codes in their definition, 39 (55%) that mentioned SNOMED codes, 15 (22%) that mentioned OMOP concept IDs, and there were 15 (22%) papers that did not mention any data standard or set of source codes. Additionally, we found that procedure codes were used in 4 (6%) phenotypes. A bar chart of source codes from the 69 publications is illustrated in Figure 1.

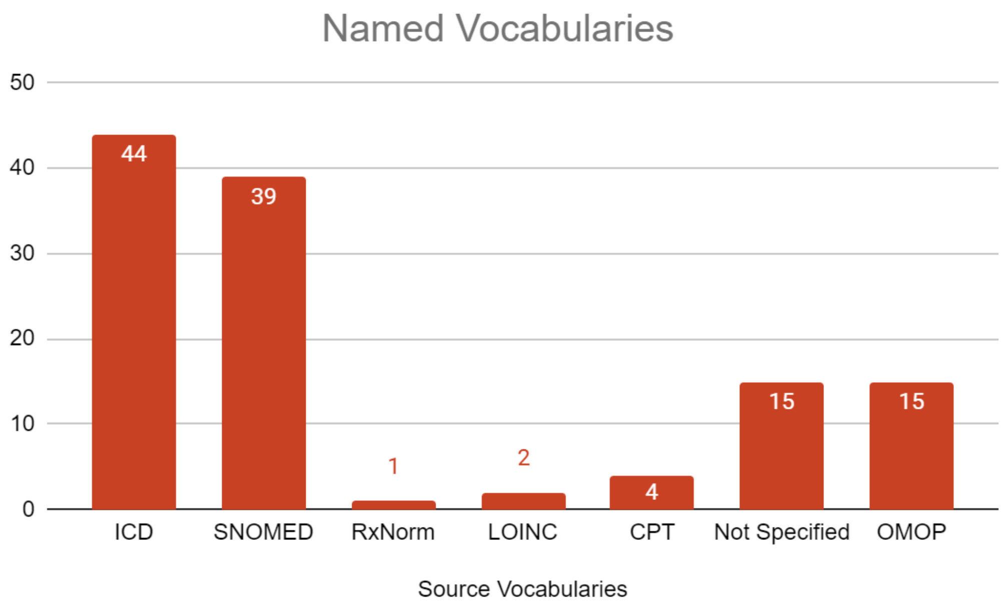

Figure 1. **Bar chart of phenotype source codes named and/or described in *All of Us* Publications.**

Conclusion

Most phenotypes used ICD diagnosis codes, a non-standard source code, which was followed closely by the OMOP standardized vocabulary, SNOMED. The RxNorm, LOINC, and CPT4 codes were used infrequently. Few studies described OMOP CDM concepts, despite the fact that the OMOP CDM is the data model for the *All of Us* Research Program. Diagnosis codes were used more than the other data domains, such as the procedure domain, regardless of the source vocabulary. The high frequency of ICD codes compared to SNOMED or OMOP concepts and diagnosis codes compared to data from other domains warrants further research. Addressing the overutilization of ICD and SNOMED concepts compared to others, the relative underutilization of procedure codes, and setting standards for phenotype definitions may improve the reliability and accuracy of observational health research.

References

1. *All of Us* data methods. National Institutes of Health *All of Us* Research Program. Accessed June 12, 2024. https://www.researchallofus.org/data-tools/methods/
2. Brandt PS, Kho A, Luo Y, et al. Characterizing variability of electronic health record-driven phenotype definitions. *J Am Med Inform Assoc*. 2023;30(3):427-437. doi:10.1093/jamia/ocac235
3. *All of Us* publications. National Institutes of Health *All of Us* Research Program. Accessed June 12, 2024. https://www.researchallofus.org/publications/?s=&post_type=publication&sort_field=publication_date&order=asc#publication-search