

Population-level characterization of individuals with whole genome sequencing samples in the All of Us Research Program

Aparajita Kashyap¹, Noémie Elhadad¹, Steven A. Kushner², Shalmali Joshi¹

¹Columbia University Department of Biomedical Informatics ²Columbia University Department of Psychiatry

Background

The All of Us (AOU) Research Program is an NIH initiative designed to collect multimodal health data from a diverse array of participants, particularly those historically underrepresented in biomedical research^{1,2}. The data may span multiple modalities, including electronic health records (EHR), health surveys, whole genome sequencing (WGS), and wearable health data (Fitbit). Through the cultivation of such datasets, multimodal research endeavors are becoming increasingly more scalable (e.g., large-scale GWAS^{3,4} and PheWAS⁵ analyses, which require both genotypic and phenotypic data). Beyond these multimodal methods, AOU WGS data has also been used for ancestry-based studies^{6,7} and rare disease analyses.⁸ Additional unimodal WGS methods, including genomic and SNP-based foundation models, have been applied to other large genomic datasets.^{9–11}

While such research can uncover valuable biological insights, typical multimodal models require that all modalities are observed. It is therefore canonical practice to leave out individuals with incomplete modalities, limiting generalizability of these representations. Incomplete observations constitute non-random, or “informative” missingness and can be driven by many factors, including patient-driven privacy concerns, sociocultural factors, and clinical considerations. We characterize population-level differences between individuals who have provided both WGS and EHR data and individuals who have only provided WGS data, thus informing hypotheses about the generalizability of work that combines EHR-derived phenotypes and genomic data beyond the WGS + EHR cohort.

Methods

We defined the WGS-only cohort as individuals with short-read WGS, but no EHR recorded in AOU, and the WGS + EHR cohort as individuals with both short-read WGS and EHR data recorded in the AOU platform. We pulled survey and zip-code data based on the most recent survey responses available and identified whether wearable data was collected for each participant. We compared the completeness of each of these data types between cohorts using a multivariate logistic regression, where we reported the adjusted odds ratio and associated 95% confidence interval. An odds ratio greater than one indicates that individuals who have completed the given data type are more likely to have both EHR and WGS data. We additionally characterized demographics (race, gender, ethnicity, and age), where age is measured as of 01/01/2026. We compared demographics between cohorts using a multivariate logistic regression as described above. We further compared the proportion of individuals with self-reported disabilities in each cohort, assessing significance with Fisher’s Exact Test (adjusted for multiple comparisons using the Bonferroni correction).

We next measured social determinants of health (SDOH) across each of the cohorts. We leveraged the participant’s 3-digit ZIP Codes, which are linked with the 2017 American Community Survey (ACS). For each participant, this returned the average value for SDOH-related factors within that geographic area. Importantly, this means that the ZIP-derived SDOH measures do not measure a person’s specific circumstances but rather reflect the average socioeconomic status of the area where the individual lives. We compared the distributions of these factors both between cohorts and externally, with respect to the average across the full American census. We augmented this analysis with self-reported information about healthcare access and literacy: we reported the proportion of individuals in each cohort who have delayed or been unable to access medical care, as well as those who have faced challenges understanding their healthcare situation (health literacy). We assessed significance with Fisher’s Exact Test (adjusted for

multiple comparisons using the Bonferroni correction). We repeated these analyses comparing the WGS-only cohort to an “all WGS” cohort (individuals with WGS data, regardless of EHR availability) and found similar results due to the 77% overlap between the WGS + EHR cohort and the all-WGS cohort (Figure 1a).

Results

Data Completeness

Individuals in the WGS-only cohort are more likely to have recorded Fitbit data (11.7% vs. 8.4%, aOR: 0.76 [0.75-0.78]) and are more likely to have completed the SDOH survey (50.7% vs. 42.3%, aOR: 0.60 [0.59-0.61]) relative to individuals in the WGS + EHR cohort. After adjusting for completion of other survey modules, participants in the EHR + WGS cohort had significantly higher odds of completing the Healthcare Access & Utilization survey compared to the WGS-only cohort (aOR: 1.41 [1.38-1.43]). Both cohorts provide ZIP Code information at similarly high rates (WGS-Only: 96.2%; WGS + EHR: 96.0%) (Table 1). This indicates that sharing of some modalities (i.e., surveys, wearables) does not necessarily indicate correlate with sharing EHR data.

Data Type	Completeness Rate: WGS-Only	Completeness Rate: WGS + EHR	Adjusted Odds Ratio
Fitbit Wearable Data	11.7%	8.4%	0.76 [0.75-0.78]
Social Determinants of Health Survey	50.7%	42.3%	0.60 [0.59-0.61]
Healthcare Access Survey	55.4%	54.6%	1.41 [1.38-1.43]
3-Digit ZIP Code + Socioeconomic Data	96.2%	96.0%	0.96 [0.92-1.00]

Table 1: Completeness rates of each of other data sources within AOU. We report the percent of participants in each cohort (WGS-only, WGS + EHR) who have shared data within each of these categories. Adjusted odds ratio is derived from multivariate logistic regression and reported alongside the 95% confidence interval.

Demographics

Individuals in the AOU Dataset follow a bimodal age distribution (Figure 1b); on average, individuals from the EHR + WGS cohort are older ($p = 0.0$). The distribution of participant gender is similar across both cohorts, with a higher prevalence of nonbinary individuals (aOR: 0.77 [0.61-0.98]) and participants who elected to not share their gender (aOR: 0.77 [0.62-0.95]) in the WGS-only cohort. Individuals in the WGS-only cohort are more likely to be Asian (aOR: 0.63 [0.50-0.78]) and Hispanic or Latino (aOR: 0.92 [0.88-0.96]), but cohorts are well matched across other racial groups (Table 2). Individuals in the EHR + WGS cohort are also significantly more likely to report having a disability across all six recorded disability categories (blind, deaf, difficulty concentrating, dressing and bathing, running errands alone, and walking/climbing). These differences are highest for individuals who report a disability related to walking/climbing (Figure 2).

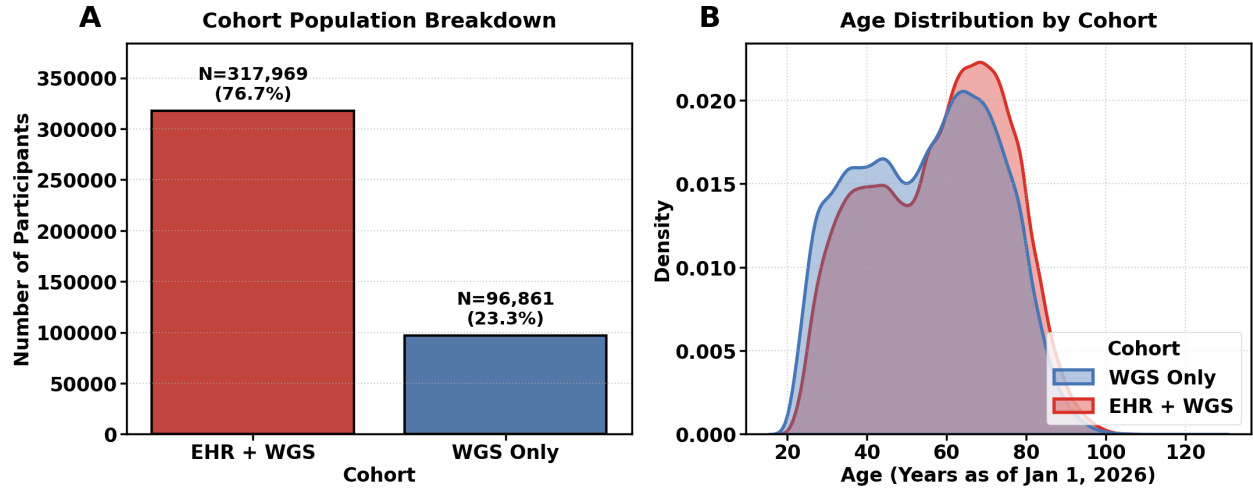


Figure 1: Cohort population. a. Of all AOU participants with WGS data, 76.7% of them also have EHR data. b. Age (years) of participants in each cohort follows a bimodal distribution, with a higher proportion of younger individuals in the WGS-only cohort.

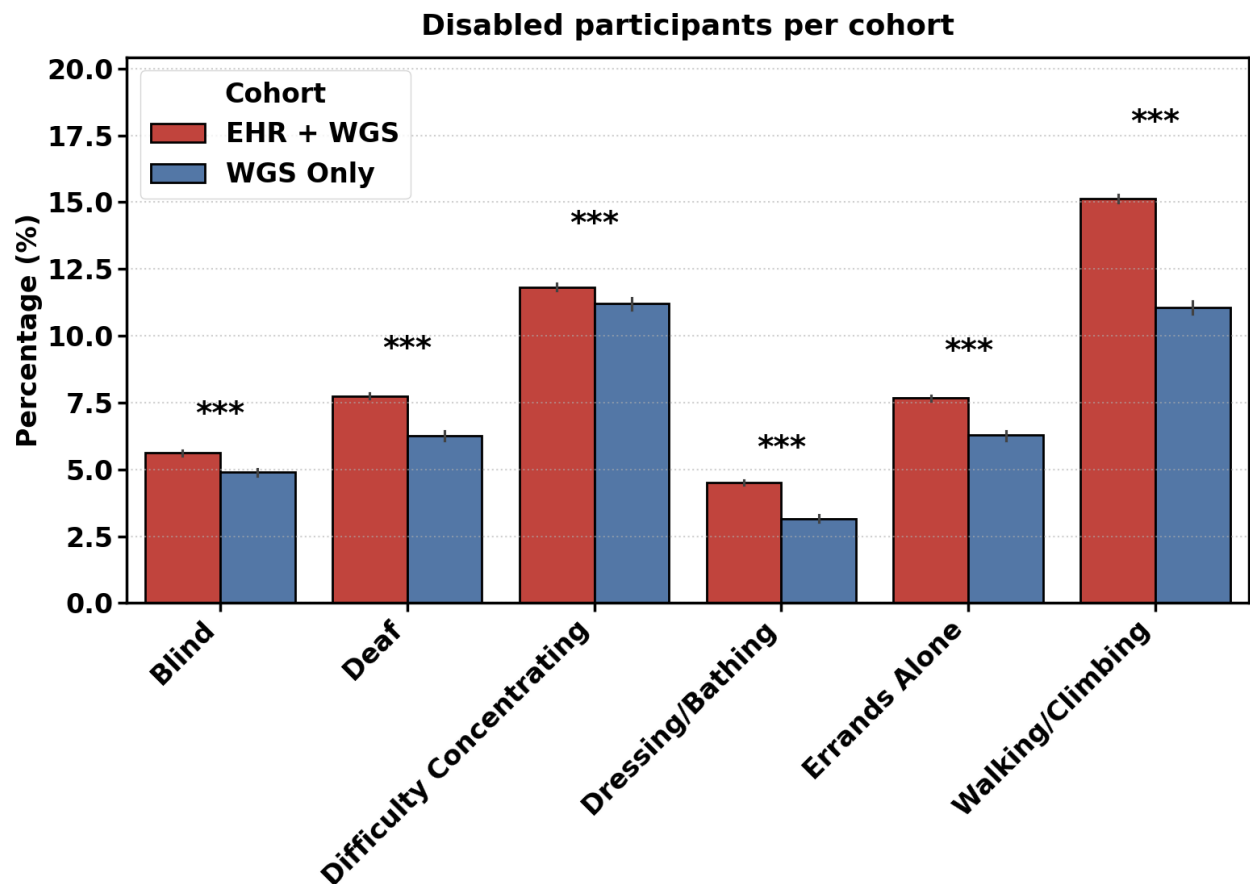


Figure 2: Participants with a self-reported disability across each cohort. We report the percent of individuals who disclose a disability in the “Basics” survey (100% completion rate) within All of Us. The

stars indicate that across all disability categories, individuals in the EHR + WGS cohort were more likely to report a disability with statistical significance $p < 0.001$ (Fisher's Exact Test).

Social determinants of health

When examining ZIP-imputed socioeconomic status (Figure 3), we find similar distributions in the rates of high school education, health insurance, vacant housing, and individuals in poverty. Interestingly, when comparing to the USA more broadly, individuals in both cohorts belong to geographic locations with higher-than-average rates of poverty and lower-than-average rates of high school completion.

Individuals in the WGS-only cohort reported slightly higher rates of healthcare access barriers: 41.1% of individuals reported being unable to afford care, while 38.8% reported experiencing delayed medical care. These numbers are 40.1% ($p = 0.016$) and 35.8% ($p = 4.12e-32$) in the WGS + EHR cohort, respectively. However, individuals in the WGS-only cohort had slightly lower rates of insurance (85.8%) relative to individuals in the WGS + EHR cohort (87.6%, $p = 1.35e-24$). Individuals in the WGS + EHR cohort also reported lower levels of health literacy (WGS-only: 26%; WGS + EHR: 29%; $p = 1.06e-15$).

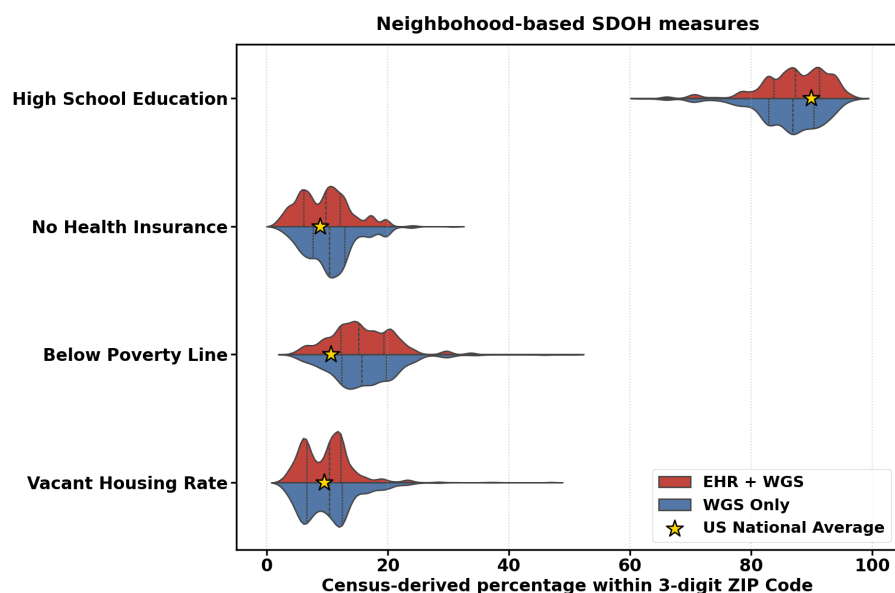


Figure 3: Distribution of area-level SDOH variables based on three-digit Zip Code data. This graph represents the distribution of area-level sociodemographic information across individuals in the WGS + EHR cohort and the WGS-only cohort. Participants were mapped to a three-digit Zip Code, and this Zip Code was linked to the 2017 American Community Survey (census) to identify area-level sociodemographic averages. The yellow stars represent the national average.

Conclusion

Individuals in the WGS-only cohort have better data completeness across non-EHR data types (i.e. optional surveys, geographic information, wearable data). This indicates that individuals in the WGS-only cohort likely do not experience disproportionate barriers to sharing non-genomic information (although there may be barriers to specifically sharing EHR data). This additionally suggests that non-EHR data could be a valuable resource for conducting GWAS or PheWAS studies that require phenotypic data.¹² Overall, individuals in the WGS-only cohort are more likely to be younger, Asian, and/or Hispanic or Latino. They are also more likely to have a disability. This indicates that based on demographics, we are unlikely to be able to generalize from the WGS + EHR cohort to the WGS-only cohort. However, we observe mixed

results for distributions of healthcare access and social determinants of health. While individuals from the WGS-only cohort experience more barriers to healthcare access, they also report lower rates of health illiteracy. This indicates that there is no clear divide in socioeconomic status and SDOH-related experiences between cohorts; rather, we must consider the specific variables or experiences we want to assess generalizability across and compare distributions within those variables.

AOU outlines the steps that participants can take to join the biobank: first individuals opt in or out of sharing their EHR; if they opt out, they are asked to fill out three surveys (Basic, Overall Health, and Lifestyle) but are not invited to provide biological samples. If they do opt in to share their EHR data, they are then asked to fill the three surveys and are then asked to provide biological samples. Additionally, other surveys and activities (e.g., geographic location, healthcare access and utilization, social determinants of health) are provided to eligible participants for optional completion. This indicates that the gap in EHR data is likely due to lack of data sharing or interoperability between the AOU database and some health care networks.

Broadly, this has two major implications: first, results generated in a combined EHR+WGS cohort may not fully generalize to the unimodal (WGS) cohort. For example, models that leverage both WGS and EHR data may not generalize to the WGS-only population due to demographic distribution shifts. This could impact supervised predictive models that leverage EHR data to generate labels (e.g., GWAS, PheWAS) or employ multimodal approaches to enrich WGS data with EHR data. Second, these results illustrate populations where future efforts can focus: by investing in data-sharing capabilities with health care networks and providers, we can enrich existing data and “fill in the gaps” of the AOU research program.

We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health’s All of Us Research Program for making available the participant data examined in this study.

		WGS-Only	WGS + EHR	Adjusted Odds Ratio
Total		96,861	317,969	
Age: Mean (SD)		54.8 (17.2)	58.0 (17.0)	1.01 [1.01-1.01]
Gender: n (%)	Female	56,594 (58.4%)	191,043 (60.10%)	1.10 [0.89-1.36]
	Male	37,841 (39.1%)	121,322 (38.20%)	1.00 [0.81-1.24]
	Non-Binary	482 (0.5%)	962 (0.3%)	0.77 [0.61-0.98]
	Transgender	187 (0.2%)	465 (0.1%)	0.90 [0.69-1.18]
	Missing/Skip*	1,638 (1.7%)	3,845 (1.20%)	0.77 [0.62-0.95]
Race: n (%)	American Indian or Alaska Native	1,134 (1.2%)	4,620 (1.50%)	1.15 [0.91-1.44]
	Asian	4,183 (4.3%)	9,307 (2.90%)	0.63 [0.50-0.78]
	Black or African American	16,283 (16.8%)	56,255 (17.70%)	0.91 [0.73-1.13]

	Middle Eastern or North African	548 (0.6%)	1,803 (0.60%)	0.91 [0.72-1.16]
	Native Hawaiian or Other Pacific Islander	100 (0.1%)	356 (0.10%)	1.15 [0.91-1.44]
	White	50,074 (51.7%)	173,923 (54.70%)	0.87 [0.69-1.08]
	More than one population	13,449 (4.2%)	13,449 (4.20%)	0.85 [0.68-1.06]
	Missing/Skip*	20,170 (20.8%)	58,256 (18.3%)	0.83 [0.66-1.04]
Ethnicity: n (%)	Hispanic or Latino	20,754 (21.4%)	58,950 (18.50%)	0.92 [0.88-0.96]
	Not Hispanic or Latino	73,373 (75.8%)	250,136 (78.70%)	0.99 [0.93-1.05]

Table 2: This table reports the demographics for each cohort (WGS + EHR; WGS-only). Adjusted odds ratio is derived from multivariate logistic regression and reported alongside the 95% confidence interval.

References

1. Bick AG, Metcalf GA, Mayo KR, Lichtenstein L, Rura S, Carroll RJ, et al. Genomic data in the All of Us Research Program. *Nature*. 2024 Mar;627(8003):340–6. doi:10.1038/s41586-023-06957-x
2. Kidenya BR, Mboowa G. Inclusiveness of the All of Us Research Program improves polygenic risk scores and fosters genomic medicine for all. *Commun Med*. 2024 Nov 8;4(1):227. doi:10.1038/s43856-024-00647-z
3. Smith K, Deutsch AJ, McGrail C, Kim H, Hsu S, Mandla R, et al. Multi-ancestry Polygenic Mechanisms of Type 2 Diabetes Elucidate Disease Processes and Clinical Heterogeneity. *medRxiv*. 2023 Sep 29;2023.09.28.23296294. doi:10.1101/2023.09.28.23296294 PubMed PMID: 37808749; PubMed Central PMCID: PMC10557820.
4. Newbury A, Jiang X, Natarajan* K, Gürsoy* G. Cross Biobank Comparison of Phenomic Profiles. *AMIA Annu Symp Proc*. 2026 Feb 14;2025:939–48. PubMed PMID: 41726399; PubMed Central PMCID: PMC12919452.
5. Khajouei E, Ghisays V, Piras IS, Martinez KL, Vicenti AT, Naymik M, et al. Phenome-wide association of APOE alleles in the All of Us Research Program. *eBioMedicine*. 2025 Jul 1;117. doi:10.1016/j.ebiom.2025.105768 PubMed PMID: 40451134.
6. Gouveia MH, Meeks KAC, Borda V, Leal TP, Kehdy FSG, Mogire R, et al. Subcontinental genetic variation in the All of Us Research Program: Implications for biomedical research. *Am J Hum Genet*. 2025 Jun 5;112(6):1286–301. doi:10.1016/j.ajhg.2025.04.012 PubMed PMID: 40480197; PubMed Central PMCID: PMC12256905.
7. Martinez KL, Klein A, Martin JR, Sampson CU, Giles JB, Beck ML, et al. Disparities in ABO blood type determination across diverse ancestries: a systematic review and validation in the All of Us Research Program. *J Am Med Inform Assoc JAMIA*. 2024 Jun 25;31(12):3022–31. doi:10.1093/jamia/ocae161 PubMed PMID: 38917427; PubMed Central PMCID: PMC11631141.

8. Updated DPYD HapB3 haplotype structure and implications for pharmacogenomic testing - PMC [Internet]. [cited 2026 May 19]. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10777430/#cts13699-sec-0004>
9. Pearce MT, Doms T, Yamamoto R, Meehl J, Molnar C, Bissell M, et al. EVEE: Interpretable variant effect prediction from genomic foundation model embeddings [Internet]. bioRxiv; 2026 [cited 2026 May 3]. p. 2026.04.10.717844. Available from: <https://www.biorxiv.org/content/10.64898/2026.04.10.717844v1> doi:10.64898/2026.04.10.717844
10. Shyr D, Dey R, Li X, Zhou H, Boerwinkle E, Buyske S, et al. Semi-supervised machine learning method for predicting homogeneous ancestry groups to assess Hardy-Weinberg equilibrium in diverse whole-genome sequencing studies. *Am J Hum Genet.* 2024 Oct 3;111(10):2129–38. doi:10.1016/j.ajhg.2024.08.018
11. Xu AG, Xu Y, Xing Y, Luo P, Yang J, Bai Y, et al. SNPBag: A Foundation Model for Multitask Genome-Scale SNP Analysis [Internet]. bioRxiv; 2025 [cited 2026 Apr 29]. p. 2025.01.29.635579. Available from: <https://www.biorxiv.org/content/10.1101/2025.01.29.635579v2> doi:10.1101/2025.01.29.635579
12. Newbury A, Elhussein A, Gürsoy G. Multi-domain rule-based phenotyping algorithms enable improved GWAS signal. *Npj Digit Med.* 2025 Aug 2;8(1):499. doi:10.1038/s41746-025-01815-8