Received: 13 October 2023 | Revised: 31 January 2024 | Accepted: 2 February 2024

DOI: 10.1002/trc2.12462

Alzheimer's & Dementia
Translational Research & Clinical Interventions

RESEARCH ARTICLE

The power of representation: Statistical analysis of diversity in US Alzheimer's disease genetics data

Diane Xue[1] | Elizabeth E. Blue[1,2,3] | Matthew P. Conomos[4] | Alison E. Fohner[1,5]

[1]Institute for Public Health Genetics, University of Washington School of Public Health, Seattle, Washington, USA

[2]Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington, USA

[3]Brotman Baty Institute, Seattle, Washington, USA

[4]Department of Biostatistics, University of Washington School of Public Health, Seattle, Washington, USA

[5]Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA

Correspondence
Diane Xue, Institute for Public Health Genetics, University of Washington School of Public Health, 1959 NE Pacific St, Room H-690, Seattle, WA 98195, USA.
Email: dxue@uw.edu

Funding information
National Institutes of Health, Grant/Award Numbers: K01 AG071689, F99 AG079792

Abstract

INTRODUCTION: Alzheimer's disease (AD) is a complex disease influenced by genetics and environment. More than 75 susceptibility loci have been linked to late-onset AD, but most of these loci were discovered in genome-wide association studies (GWAS) exclusive to non-Hispanic White individuals. There are wide disparities in AD risk across racially stratified groups, and while these disparities are not due to genetic differences, underrepresentation in genetic research can further exacerbate and contribute to their persistence. We investigated the racial/ethnic representation of participants in United States (US)-based AD genetics and the statistical implications of current representation.

METHODS: We compared racial/ethnic data of participants from array and sequencing studies in US AD genetics databases, including National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) and NIAGADS Data Sharing Service (dssNIAGADS), to AD and related dementia (ADRD) prevalence and mortality. We then simulated the statistical power of these datasets to identify risk variants from non-White populations.

RESULTS: There is insufficient statistical power (probability <80%) to detect single nucleotide polymorphisms (SNPs) with low to moderate effect sizes (odds ratio [OR]<1.5) using array data from Black and Hispanic participants; studies of Asian participants are not powered to detect variants OR <= 2. Using available and projected sequencing data from Black and Hispanic participants, risk variants with OR = 1.2 are detectable at high allele frequencies. Sample sizes remain insufficiently powered to detect these variants in Asian populations.

DISCUSSION: AD genetics datasets are largely representative of US ADRD burden. However, there is a wide discrepancy between proportional representation and statistically meaningful representation. Most variation identified in GWAS of non-Hispanic White individuals have low to moderate effects. Comparable risk variants in non-White populations are not detectable given current sample sizes, which could lead to disparities in future studies and drug development. We urge AD genetics

researchers and institutions to continue investing in recruiting diverse participants and use community-based participatory research practices.

KEYWORDS
Alzheimer's disease, ancestry, community-based participatory research, diversity, ethnicity, genetics, race, representation, statistical power

1 | INTRODUCTION

By 2050, >12 million people in the United States will have Alzheimer's disease (AD), and the risk for AD is not evenly distributed across the population.[1] A recent study of age-adjusted incidence among 1.8 million veterans found substantially higher rates of dementia among self-reported Hispanic and Black participants compared to American Indian or Alaska Native (AI/AN) participants, with the lowest incidence among Asian and White participants: estimates ranged from 20.7, 19.4, 14.2, 12.4, and 11.5 per 1000 person-years, respectively.[2] While AD and dementia are not synonymous, AD is the most common cause of dementia, accounting for 60%–80% of dementia cases.[3] These results are consistent with previous studies that have shown disparities in AD outcomes across racialized groups.[1] Much of these differences arise from inequalities in social determinants, such as those influencing education and risk for cardiovascular disease and hypertension, which are known risk factors for AD and dementia.[4,5]

AD risk is also strongly influenced by genetics. AD is a complex, highly heritable (h^2 = 58%–79%) disease.[6] To date, >75 susceptibility loci have been implicated in late-onset AD.[7,8] Due to differences in linkage disequilibrium and both genetic and non-genetic modifiers, the genetic architecture of AD differs across ancestry groups in terms of associated variants and effect sizes of commonly implicated variants.[9,10–12] Because demographic histories create structure in human genetics,[13,14] differences in allele frequency and linkage disequilibrium across global populations correlate with racialized groups.[11] While wide disparities in AD risk across racially stratified groups are not caused by genetic differences, inequality in genetic research can further exacerbate health disparities and contributes to their persistence. Most known risk variants were discovered in genome-wide association studies (GWAS), which now include >1 million participants, primarily focused on self-described non-Hispanic White individuals who cluster with 1000 Genomes Project (1KG) European ancestry groups (EUR).[7,17] Meanwhile, the largest GWAS of African and African American individuals that cluster with the African Genome Resources Reference (AA) included a mere 2784 cases and 5222 controls[16]—and studies include even fewer participants for other populations. This disparity translates to an understanding of AD genetic architecture that is both incomplete and inequitable.[16–18]

Because this paper focuses on representation in genetic studies, it is important to distinguish between biological and social population descriptors.[19] Race/ethnicity are socially constructed without biological meaning, while genetic ancestry refers to the continental or geographic origins of biological ancestors.[11] Our study relies on population descriptors aligned with social categorizations for both practicality and future use of findings. Our study is based on previous data collection efforts, and the demographics reported in previous studies are typically socio-political categorizations in adherence with US Office of Management and Budget (OMB) standards.[20] Furthermore, because genetic ancestry is not known at the time of recruitment, and barriers and willingness to participate in genetic research are more closely related to social and environmental differences, using social categorization when describing participants is more relevant for future applications.[21] When the studies contributing to our work describe procedures of using genetic ancestry information to filter participants, such as using principal components, we will describe the genetic reference used for filtering in addition to the self-reported or ascribed racial/ethnic categorization used for recruitment (i.e., EUR = non-Hispanic White individuals who cluster with 1KG European ancestry groups).

Over a decade since the launch of the National Alzheimer's Project Act, we are on the cusp of its initial goal to prevent and effectively treat AD by 2025.[22] Now is a critical time to assess the state of representation and diversity in AD genetics research. The National Institutes of Health (NIH) allocates >$3 billion annually to deepen our understanding of AD and facilitate the development of effective treatments. It is crucial, however, to ensure equity in who is benefiting from this extensive investment. The NIH has devoted resources to this effort, including the launch of Outreach Pro (https://outreachpro.nia.nih.gov/), which provides study recruitment materials in multiple languages and funding the Alzheimer's Disease Sequencing Project (ADSP) Follow-Up Study 2.0 Diversity Initiative Phase, which is committed to identifying therapeutic targets benefitting a diverse population.[23] Here, we investigate how well US-based AD genetic datasets represent the racial and ethnic demographic characteristics of those living with AD in the United States, and whether current and planned AD genetics studies are adequately powered to advance racial/ethnic equity in our understanding of the genetic architecture of AD. We conclude by offering suggestions for future recruitment priorities for AD genetics studies.

2 | METHODS

2.1 | Quantifying AD burden in the United States

We aimed to quantify the demographics of AD burden in the United States by estimating disease prevalence by race and ethnicity. We categorized individuals by self-reported or ascribed race/ethnicity into five

RESEARCH IN CONTEXT

1. **Systematic review**: The authors collected race/ethnicities of participants in United States (US)-based genetic studies of Alzheimer's disease (AD) available on the National Institute for Aging Genetics of Alzheimer's Disease Storage Site (NIAGADS).
2. **Interpretation**: While racial/ethnic demographics of US Alzheimer's genetics studies largely reflect the population living with AD, proportional representation is not equitable. Most variants detected thus far in studies of non-Hispanic White individuals with 1KG-European ancestry are of low effect size. Despite efforts to increase diversity in AD sequencing datasets, sample sizes remain insufficient to detect comparable genetic variation in populations genetically distant from 1KG-European ancestry. As genetic research informs downstream epidemiological research, drug discovery, and disease prediction, varied genetic knowledge may exacerbate existing disparities in AD prevention, diagnosis, and treatment.
3. **Future directions**: The current understanding of AD genetics is inequitable. AD genomics researchers must embrace community-based participatory research strategies to build trust and avoid perpetuating disparities throughout the research pipeline.

groups defined by the US OMB,[20] which guides how the federal government collects ethno-racial data: AI/AN, Asian, Black, Hispanic or Latino, and White. Participants who identified as "other" were excluded from the analysis.

The most widely reported AD prevalence estimates are based on forward projections derived from the Chicago Health and Aging Project (CHAP).[24] This study estimated prevalence for non-Hispanic White, Black, and Hispanic individuals but did not estimate prevalence for Asian or AI/AN group. To approximate the AD burden in Asian and AI/AN people, we used estimates of dementia prevalence for AI/AN, Asian, Black, Hispanic, and White individuals based on Medicare Fee-for-Service beneficiaries and the US Census data.[25]

Because dementia prevalence includes non-AD dementia, we analyzed AD mortality as a supplemental measure of the public health burden. We obtained de-identified age-adjusted mortality data for AD in the United States from the Centers for Disease Control Wide-Ranging Online Data for Epidemiologic Research (CDC WONDER) Underlying Cause of Death database.[26] CDC WONDER data are based on death certificates for US residents, collected from 1999 to 2020. This dataset considers one underlying cause of death per person. Deaths for 1999 and beyond are classified using the Tenth Revision of the International Classification of Disease (ICD). Race and ethnicity are obtained either from self-report prior to death or reported by surviving next of kin, an informant, or by observation. We queried crude and age-adjusted death rates due to AD by race, Hispanic ethnicity, and year for the most recent five years of data availability (2016–2020) using the same five racial/ethnic categories defined by the OMB guidelines. Crude proportions can be compared to the other sources of AD population demographic data, but because racial and ethnic groups represent different proportions of the US population and have different average age-at-death, we also evaluated age-adjusted mortality rates.

Comparison between proportions of disease burden from the three sources was performed using a chi-squared test for proportions.

2.2 | Quantifying racial/ethnic representation in genetic datasets

We obtained demographic data for participants in US AD genetic studies within the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, https://www.niagads.org/) and the NIAGADS Data Sharing Service (dssNIAGADS, https://dss.niagads.org/). NIAGADS is responsible for harmonizing and sharing AD genetics, genomics, and phenotypic data derived from NIA-funded AD genetics studies. Access to this publicly available data was approved by NIAGADS. We reviewed all genotype array datasets within NIAGADS that met the following criteria: Disease = "AD," Molecular Data = "Genotype," Type = "GWAS." Array datasets from dssNIAGADS were selected by filtering for Disease = "AD" and Data Type = "GWAS". Some AD GWAS data predate NIAGADS and are stored elsewhere. We therefore also include the following US-based AD GWAS data sets with clinical phenotyping and race/ethnicity data not captured in NIAGADS: African American Alzheimer's Disease Genetics Study, Alzheimer's Disease Neuroimaging Initiative, BIOCARD, Cohorts for Heart and Aging Research in Genomic Epidemiology consortium (CHARGE; includes Atherosclerosis Risk in Communities Study, Cardiovascular Health Study, and Framingham Heart Study), and the Genetic and Environmental Risk Factors for Alzheimer Disease Among African Americans Study. All participant demographic data for AD sequencing studies are from the Alzheimer's Disease Sequencing Project (ADSP) Umbrella Study, obtained from dssNIAGADS, representing 25 datasets. In addition to sequencing data that are currently available, we analyzed the reported demographics of whole genome sequencing data that ADSP has planned for release through 2027.

Race/ethnicity data were extracted directly from study-specific covariate files where possible, or either approximated from study-specific publications or obtained directly through correspondence with study coordinators. Ancestry, race, and ethnicity labels have been inconsistently used across AD GWAS. The largest GWAS with individuals who identify as White uses "European ancestry" as inclusion criteria,[15,17,18,27] while the largest GWAS with individuals who identify as Black is referred to as a study of "African American" individuals.[16] We therefore grouped labels when necessary, for example, "Caucasian" was grouped with "White," "African American" with "Black." All Hispanic participants were evaluated exclusively as Hispanic and were not included in a racial category (i.e., White = non-Hispanic White).

We evaluated participant demographics separately for array, whole-exome (WES), and whole-genome sequencing (WGS) data, as these could be considered different types of biological data. Array-based data are restricted to a pre-selected array of single nucleotide polymorphisms (SNPs). Sequencing data are a read-out of every base-pair in one's exome or genome. Array data were more popular historically due to the relative ease and lower cost of conducting the assays, but more recent studies have favored WES and WGS data as costs have decreased and technology has improved. Some individuals represented in array data are represented in sequencing data.

Chi-squared tests for proportions were used to compare disease burden in the population and racial/ethnic representation in AD genetics datasets.

2.3 | Determining statistical power of existing and planned data

We conducted power analyses for hypothetical GWAS of AD case-control status stratified by race based on demographics of participants across all available datasets. GWAS continue to be the dominant method used to identify risk alleles in populations. Power was simulated separately for array and sequencing data using the R function *genpwr::genpwer.calc* (version 1.0.4), available on CRAN. We assumed an additive model and simulated case rates based on population-specific case proportions of each dataset. More information on study case proportions and genpwr case rate selection are included in the supplement (Tables S1-S3). We simulated power to identify variants with odds ratios (ORs) equal to 1.1, 1.2, 1.5, and 2 given significance levels of $p <$ 5e-08 (genome-wide) and $p <$ 2.5e-06 (exome-wide), and a continuous range of minor allele frequencies from 0 to 0.5. We define "low" effect size as OR = 1.1 or less (ex., *ACE*[17]), "modest" effect size as OR = 1.2 (ex., *BIN1*[28]), "intermediate" as OR = 1.5 (ex., *NCK2*[29]), and "high" effect size as OR = 2 (ex., *TREM2*[30,31]) or more. These designations follow the most recent comprehensive review of the genetic architecture of AD.[7]

3 | RESULTS

3.1 | Quantifying AD burden in the United States

Approximately 6.7 million adults aged 65 and older are currently living with AD in the United States with the following distribution across racial and ethnic groups: 70.8% White, 17.4% Black, and 11.7% Hispanic (Table 1). Because Asian and AI/AN individuals were not represented in CHAP, we extended our analyses to dementia prevalence values,[25] the majority of which represent AD (60%–80%[3]). Dementia prevalence estimates were consistent with the AD prevalence estimates, where those categorized as White individuals made up most of projected dementia cases (72.7%), followed by Black (12.6%), Hispanic (10.3%), Asian (3.7%), and AI/AN (0.6%) individuals (Table 1). United States cause-of-death estimates from the CDC WONDER database indicate 83.4% of AD deaths were among individuals identified as White, a slightly higher proportion than our AD and dementia prevalence estimates, followed by Black (7.6%), Hispanic (6.4%), Asian (2.4%), and American Indian/Alaskan Native (0.3%) individuals.

Racial and ethnic groups represent different proportions of the US population and have different average age-at-death, which can be accounted for in age-adjusted mortality rates. AD mortality rates per 100,000 individuals were as follows: White (254.2), Black (223.9), Hispanic (213.7), AI/AN (151.8), and Asian (125.6) (Table S4) individuals. The proportional representation of racial and ethnic groups across AD prevalence, dementia prevalence, and AD mortality did not significantly differ ($X^2 = 10.711$, $p = 0.2186$, Bonferroni-corrected $\alpha = 0.0167$).

3.2 | Quantifying proportional representation of existing genetic data

AD GWAS studies using array data are proportionally representative of AD (Figure 1, Table 1). We identified 36 genotype array datasets encompassing 65,733 individuals (Table S1); among them, 77% of participants are classified as White, 14.4% Black, 6.8% Hispanic, 1.8% Asian, and 0.02% AI/AN. These proportions are similar to those in our AD and dementia prevalence estimates above, and do not differ significantly (Table 2).

The currently available sequencing data are more diverse than the array data, mostly due to better representation of Hispanic populations. Figure 1 displays 17 available WGS sample sets that are part of the ADSP Umbrella encompassing 36,336 individuals (Table S2); among these participants, 45.0% are classified as White, 15.7% Black, and 31.1% Hispanic, while Asian and AI/AN participation remains low (7.8% and 0.4%, respectively.) Almost all Asian participants with genetic sequencing are from the Harmonized Diagnostic Assessment of Dementia for the Longitudinal Aging Study of India (LASI-DAD) study, a subset of the Longitudinal Aging Study in India. Unlike other included studies, LASI-DAD participants are not from the United States, but the study is funded and administered by US institutions and investigators, and data are stored in US repositories. While the inclusion of LASI-DAD is a significant improvement for South Asian representation compared to the array data, there is little improvement in representation for East and Southeast Asians. The proportion of each racial/ethnic group in the WGS studies significantly differed from the racial/ethnic proportions of AD prevalence, dementia prevalence, and AD mortality (Table 2).

3.3 | Determining statistical power of existing genetic data

We conducted power calculations using the sample sizes derived from existing array and sequencing data as well as planned WGS data releases to ascertain the ability to identify association signals in GWAS stratified by race/ethnicity. Power calculations simulated genotype array data using the following sample sizes: 50,000 non-Hispanic

TABLE 1 Population-specific AD burden and representation in AD genetics data in the United States.

Race/ethnicity	Projected AD prevalence 2020 in 1000s (%)	Projected dementia prevalence 2020 in 1000s (%)	Deaths 2016-2020 (%)	US population (2000 standard)	Array data sample size (%)	WGS data sample size (%)
AI/AN	–	38 (0.6)	1865 (0.3)	1,584,958	14 (0.02)	152 (0.42)
Asian/PI	–	212 (3.7)	14,272 (2.4)	12,526,017	1170 (1.78)	2820 (7.76)
Black	1060 (17.4)	726 (12.6)	45,946 (7.6)	24,282,298	9439 (14.36)	5695 (15.67)
Hispanic	710 (11.7)	594 (10.3)	38,960 (6.4)	20,361,950	4491 (6.83)	11,329 (31.18)
White	4300 (70.8)	4186 (72.7)	505,889 (83.4)	201,746,665	50,619 (77.01)	16,340 (44.97)

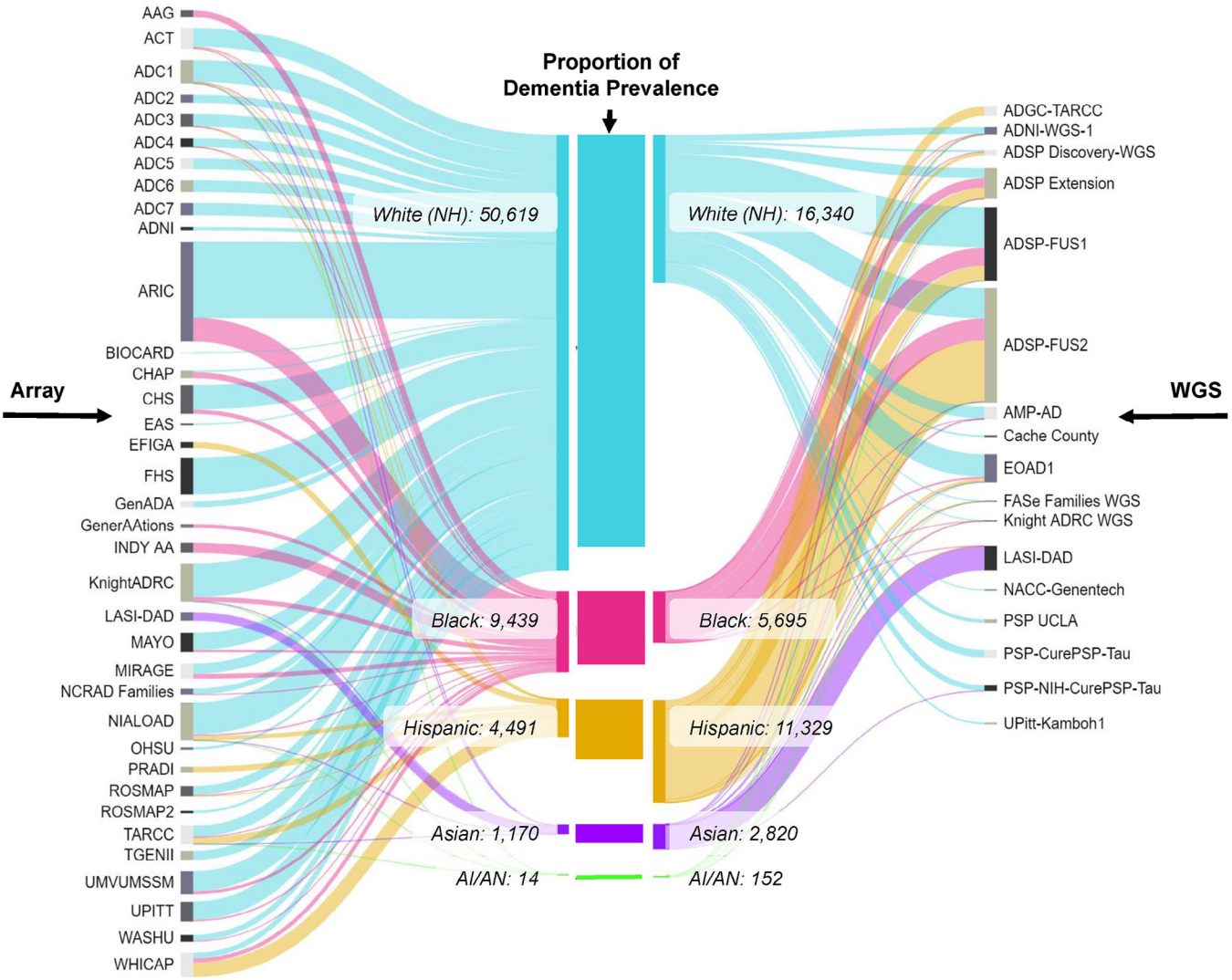

FIGURE 1 Racial/ethnic profile of AD genetic data. The left-hand side depicts a Sankey plot showing racial/ethnic representation in each array dataset, flowing from left to right. The right-hand side depicts a Sankey plot of racial/ethnic representation in the whole genome sequencing (WGS) data, flowing right to left. On the outside edges are the individual cohorts. The participants are grouped by five broad racial/ethnic categories: White (non-Hispanic), Black, Hispanic, Asian, and American Indian/Alaska Native (AI/AN). In the center of the figure is the relative race/ethnicity specific burden of dementia. Exact proportions of dementia burden are found in Table 1.

TABLE 2 Statistical comparison of observed race/ethnicity representation in AD genetics studies versus expected values based on AD and dementia prevalence and mortality.

Parameter	Array	Sequencing
AD prevalence	$X^2 = 3.63, p = 0.458$	$X^2 = 22.88, p = 0.0001$*
Dementia prevalence	$X^2 = 2.16, p = 0.707$	$X^2 = 18.849, p = 0.0008$*
AD mortality	$X^2 = 2.69, p = 0.476$	$X^2 = 33.49, p = 9.5e\text{-}07$*

*Bonferroni adjusted $\alpha = 0.00833$.

White, 8600 Black, 1800 Hispanic, and 1200 Asian participants, while calculations for currently available WGS sequencing data simulated data for 16,300 non-Hispanic White, 5700 Black, 11,300 Hispanic, and 2800 Asian participants. WGS available thru 2027 is projected to be approximately 27,000 non-Hispanic White, 18,400 Black, 29,800 Hispanic, and 7600 Asian participants. These estimated sample sizes represent a best-case-scenario assuming all samples meet quality control standards and are not duplicated across sample sets within array and sequencing data. Power simulations of WES data used the following sample sizes: 13,500 non-Hispanic White, 4400 Black, and 2200 Hispanic participants. Available genetic data for AI/AN were too small to identify any genome-wide significant hits using either genotyping array or sequencing data of any frequency or effect size ($p <$ 5E-08). Similarly, we were unable to model power to detect exome-wide significant hits ($p <$ 2.5E-06) with current sample sizes of Asian participants.

Based on sample sizes of existing genotyping array data, only studies of non-Hispanic White individuals have adequate sample sizes to detect variants with low effect sizes at genome-wide significant or suggestive thresholds (Figure 2, Figure S1, and Table S5). Sample sizes comparable to existing array data alone from Black and Hispanic individuals have insufficient statistical power (Pr [$p <$ 5E-08] <80%) to detect variants with low effect size (OR = 1.1), even when these variants are very common (frequency ~ 0.5). In the case of current array sample sizes of Hispanic participants, statistical power is only adequate to identify common variants with high effect sizes (OR ≥ 2).

WGS and WES samples remain smaller than available array data, leading to studies that remain underpowered to detect variants of low or moderate effect sizes in studies of Black and Hispanic participants (Figure 2, Figure S2, and Table S5). However, the statistical power will certainly improve as sequencing data from Black and Hispanic participants are projected to dramatically increase in the next five years. While sequencing data from Asian individuals will more than double in the next five years, studies will remain underpowered to detect common variants of low or even moderate effects.

4 | DISCUSSION

There is a wide discrepancy between proportional representation and statistically meaningful representation in AD genetic datasets. While racial/ethnic representation in older array datasets are largely comparable to proportions of AD burden in the United States, proportional sampling results in inherently unequal understanding of genetic architecture across populations—evident in the striking lack of statistical power to find genetic variants with modest effects on AD risk using all available data from non-White populations. Participation in AD sequencing studies is poised to be enriched for individuals from historically underrepresented groups relative to their proportions in AD epidemiological data. The "oversampling" is justified and necessary—these recruitment efforts have substantially increased the power of GWAS to identify AD variants with modest to intermediate effect sizes in Black and Hispanic populations. Similar population-specific breakthroughs in Asian and AI/AN population will lag as sample sizes remain insufficient for comparable discoveries. Notably, most variants identified thus far in GWAS of EUR populations have low effect sizes, and comparable discoveries in other populations continue to be unidentifiable with current sample sizes.

While most risk variants are not exclusive to any one ancestry background—AD associated SNPs first discovered in GWAS of EUR populations have been identified in other populations and vice versa (Table S6)[32,33]—gaps in statistical power continue to undermine our overall understanding of disease. Studying genomes with diverse ancestry is necessary for the discovery of novel risk variants; genomes with AA ancestry capture much more genetic diversity, with significant variation that is not present in the EUR genomes.[34] Indeed, association studies conducted in Caribbean Hispanic and African American individuals with 1KG-YRI-like variation have identified common variants in *FBXL7* and *ABCA7* not replicated in EUR due to differences in allele frequency.[35,36]

Disparities in genetic knowledge have implications for downstream applications including risk prediction and understanding underlying disease biology, drug development, and elucidating causal relationships between non-genetic risk factors and AD risk. Numerous papers have described disparities in predictive performance across diverse populations when using genetic risk prediction models developed using summary statistics from GWAS of EUR individuals.[37,38] This can result in inequalities in the ability to accurately identify individuals at high risk of disease for risk stratification in clinical trials or interventions.

There is not a one-size fits all approach for recruiting diverse participants. For example, while mistrust of biomedical research resulting from historical events (e.g., Tuskegee Syphilis Study, HeLa cells)[39] are often cited for low participation among Black and AI/AN peoples, recent studies have shown that low invitation rates may be to blame for low participation among Black individuals.[40] Meanwhile, American Indian/Alaska Native communities report a lack of involvement in study planning and use of research methods that do not respect community traditions, leading to hesitancy about participating in genomics research.[41,42] Furthermore, Asian and Hispanic participants have identified language and cultural barriers in study materials and communication as hindering their participation in genetic studies. For example, the use of the Spanish word *demencia* in study materials can dissuade participation of Hispanic participants because the meaning of *demencia* is close to "crazy."[43] Thus, efforts to increase enrollment of participants must be tailored to the target populations and their specific concerns.

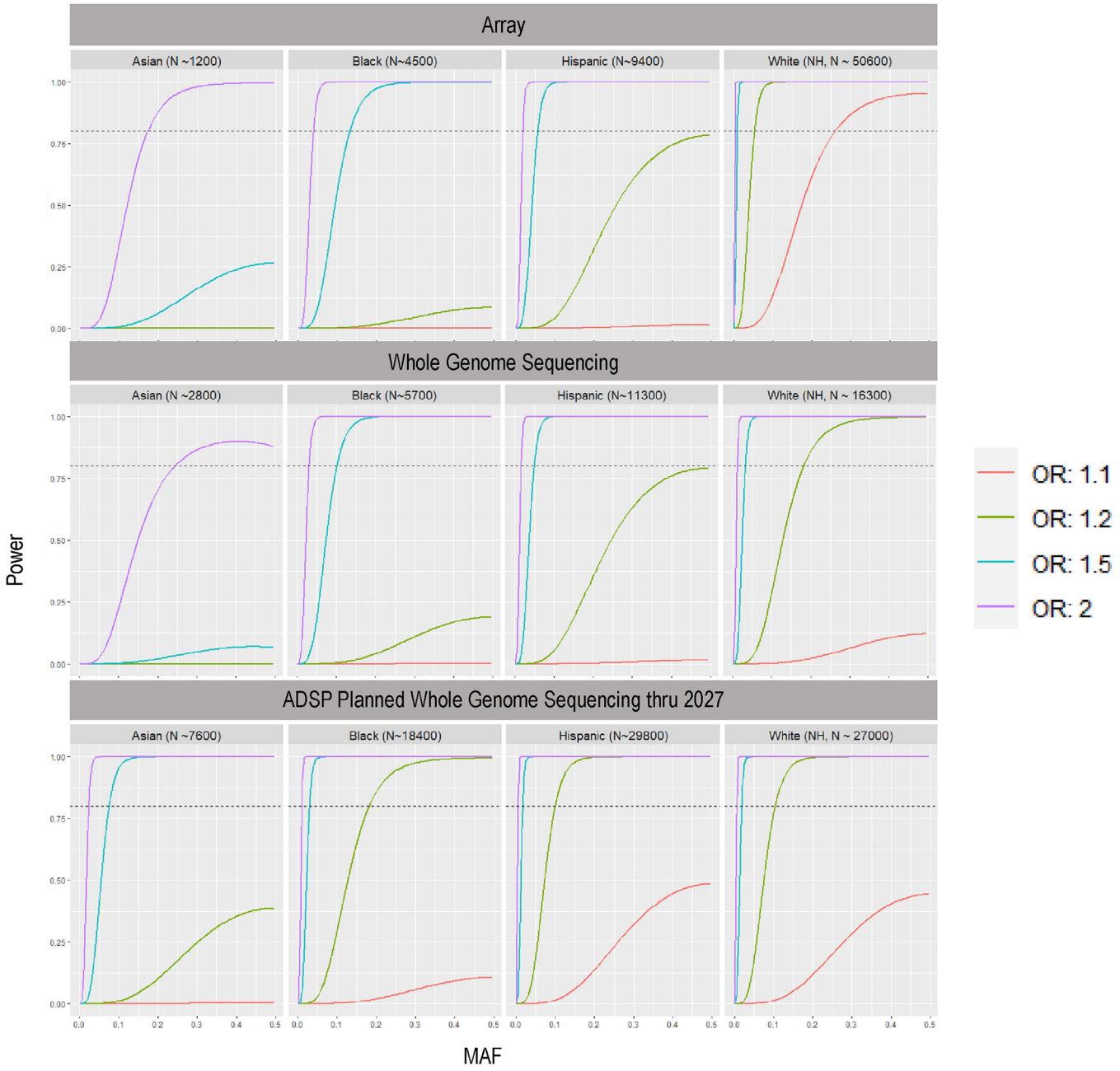

FIGURE 2 Statistical power to detect loci representing AD genetic architecture across populations. Each panel shows power to detect a significantly associated single nucleotide polymorphism (SNP) ($p = 5e\text{-}08$) for a different racial/ethnic group using genome-wide association studies (GWAS). Power simulations are based on current or projected sample sizes and case proportions. Power was simulated based on a set of effect sizes (odds ratio = 1.1, 1.2, 1.5, 2) and minor allele frequencies (MAF) ranging from 0.001 to 0.5. Separate simulations were conducted for array, current whole genome sequencing (WGS), and projected WGS data. The dashed line represents power = 0.80. American Indian/Alaska Native (AI/AN) are not included because current sample sizes are too small to detect any SNPs regardless of MAF or effect size.

All efforts to classify participants by race and/or ethnicity create large, heterogeneous, and imprecise groups. Racial/ethnic categorizations are poor proxies of environmental factors, and there are myriad socio-cultural and environmental differences within racial/ethnic groups that impact recruitment and participation.[21] One way to better address these diverse concerns is through community-based participatory research (CBPR). CBPR engages community stakeholders as peers in all stages of the study from design to dissemination of results. For example, hiring research specialists from a community to translate study materials increases the chances of using appropriate, non-stigmatizing language.[44] CBPR, though underutilized, has successfully led to increased participation of non-White populations in genetic research and could be a useful approach for increasing recruitment across many diverse populations.[45–47] Efforts for recruiting historically underserved participants into AD genetics studies using community-based approaches are underway. The Asian Cohort for

Alzheimer's Disease (ACAD), for example, is currently recruiting Asian American and Asian Canadian participants using CBPR approaches including partnership with clinics and senior homes that serve Asian communities and translation of materials into Mandarin, Cantonese, Vietnamese, and Korean.[48]

ACAD and other efforts to recruit Asian Americans are critically needed. To increase knowledge of AD genetics in Asian persons, the ADSP is primarily relying on partnerships with foreign-based studies in India and Korea.[49–51] While this strategy may help overcome potential cultural barriers to obtaining genetic data that represent individuals with genetic ancestry similar to those currently residing in India and Korea, there are limitations to interpretation and generalizability of findings conducted in these studies. First, individuals with Indian or Korean ancestry make up only a quarter of those who identify as Asian American.[52] Perhaps, more importantly to AD genetic research, GWAS associations are influenced by context,[53] and there are vast differences in environmental factors across countries that could modify genetic effects. There may also be barriers in future efforts to include social determinants and electronic health record phenotypes across AD sequencing participants that could lead to further exacerbating the disparity in AD knowledge for Asians in the United States.

Advances in statistical methods offer additional tools for increasing genetic discoveries in diverse populations. Specific ancestry backgrounds enable alternative or complementary gene-discovery approaches. For example, admixture mapping, which leverages the mix of pre-diaspora ancestry in contemporary populations, can have more statistical power than GWAS to discover genomic regions associated with traits or diseases.[54] Admixture mapping has already implicated novel AD loci in studies of African American individuals with HGDP-African/European-like ancestry and Caribbean Hispanic individuals with 1KG-CEU/YRI-like and HGDP-Pima/Maya/Colombia ancestry, despite samples sizes that are relatively small (~10,000).[55–57] Methods have been developed that allow meta-analysis of GWAS across ancestries or inclusion of participants with diverse ancestries in the same GWAS.[58] These "cross-population" GWAS may be a superior alternative to stratified studies because of the boost in statistical power from variants that are found in many populations and the reinforcement of the fact that global genetic ancestry cannot be accurately stratified into biologically meaningful "racial" groups. Indeed, the association of the *SHARPIN* genetic region with AD risk was observed in a recent multi-ancestry GWAS meta-analysis including 56,241 individuals; previously, this region had only been detected in an AD GWAS of exclusively non-Hispanic White individuals with sample size 13X greater.[18,59] While statistical advancements offer significant benefits, they do not alleviate the burden of recruiting larger sample sizes of diverse participants.

Our study has several limitations. Population-based prevalence estimates may be biased—likely underestimating the burden on Indigenous individuals and those racialized as Black who are less likely to be formally diagnosed with AD due to inequitable treatment in healthcare settings and diagnostic thresholds primarily based on White individuals.[60–62] Furthermore, the CDC WONDER dataset provides statistics for a single underlying cause of death for each person. It is likely that many people were not identified as dying with AD if another condition was a more immediate cause of their death (ex., someone living with AD who died from heart failure, in which case AD would not be listed as the underlying cause of death on the death certificate). Disparities driven by structural racism exist for chronic diseases including cardiovascular disease and cancer,[63,64] which can lead to biased underreporting of AD mortality as these causes of death may be disproportionately masking AD-related deaths. Differences in survival rates after dementia diagnosis could also contribute to the differences in proportions between prevalence and mortality rates.[65] There may also be some unreliable reporting of racial/ethnic classification of mortality data due to reporting by an observer rather than self-report. Our summaries of existing AD genetic datasets do not account for the possibility of unaccounted sample overlap within array datasets or within sequencing datasets, which would cause us to have over-estimated existing sample sizes. In this case, the problems we described would only be more important to address. Last, power calculations did not specifically model rare variant aggregate tests. Despite these constraints, we describe the "best case" of genetic data representation, which indicated that, while participation in AD genetics datasets is approximately proportional to AD burden, studies remain underpowered to elucidate the genetic architecture of AD in diverse populations.

In conclusion, we must recognize that non-White populations are simultaneously overexposed to AD risk and underrepresented in AD genetics research. Substantial effort must continue to be made to build trust, foster engagement, and actively involve historically underrepresented groups in AD genetics research to ensure that research outcomes and resulting therapies are effective for individuals of all backgrounds.

ACKNOWLEDGMENTS

We thank the UW Alzheimer's Disease Research Center for providing feedback on this work. This publication was supported by the National Institutes of Health [K01 AG071689 & F99 AG079792].

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest. Author disclosures are available in the supporting information.

CONSENT STATEMENT

Consent from human subjects was not necessary for this study.

ORCID

Diane Xue https://orcid.org/0000-0001-7633-2253

REFERENCES

1. 2023 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2023;19:1598-1695. doi:10.1002/alz.13016
2. Kornblith E, Bahorik A, Boscardin WJ, Xia F, Barnes DE, Yaffe K. Association of race and ethnicity with incidence of dementia among older adults. *JAMA*. 2022;327:1488-1495. doi:10.1001/jama.2022.3550
3. Sosa-Ortiz AL, Acosta-Castillo I, Prince MJ. Epidemiology of dementias and Alzheimer's disease. *Arch Med Res*. 2012;43:600-608. doi:10.1016/j.arcmed.2012.11.003

4. Krishnamurthy S, Rollin FG. We must be clear that the root cause of racial disparities in Alzheimer's disease is racism. *Alzheimers Dement*. 2023;19:5305-5306. doi:10.1002/alz.13389
5. Adkins-Jackson PB, George KM, Besser LM, et al. The structural and social determinants of Alzheimer's disease related dementias. *Alzheimers Dement*. 2023;19:3171-3185. doi:10.1002/alz.13027
6. Gatz M, Reynolds CA, Fratiglioni L, et al. Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry*. 2006;63:168-174. doi:10.1001/archpsyc.63.2.168
7. Andrews SJ, Renton AE, Fulton-Howard B, Podlesny-Drabiniok A, Marcora E, Goate AM. The complex genetic architecture of Alzheimer's disease: novel insights and future directions. *EBioMedicine*. 2023;90:104511. doi:10.1016/j.ebiom.2023.104511
8. Kamboh MI. Genomics and functional genomics of Alzheimer's disease. *Neurotherapeutics*. 2022;19:152-172. doi:10.1007/s13311-021-01152-0
9. Blue EE, Horimoto ARVR, Mukherjee S, Wijsman EM, Thornton TA. Local ancestry at APOE modifies Alzheimer's disease risk in Caribbean Hispanics. *Alzheimers Dement*. 2019;15:1524-1532. doi:10.1016/j.jalz.2019.07.016
10. Reitz C, Mayeux R. Genetics of Alzheimer's disease in Caribbean Hispanic and African American populations. *Biol Psychiatry*. 2014;75:534-541. doi:10.1016/J.BIOPSYCH.2013.06.003
11. Khan AT, Gogarten SM, McHugh CP, et al. Recommendations on the use and reporting of race, ethnicity, and ancestry in genetic research: experiences from the NHLBI TOPMed program. *Cell Genomics*. 2022;2:100155. doi:10.1016/j.xgen.2022.100155
12. Tishkoff SA, Verrelli BC. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet*. 2003;4:293-340. doi:10.1146/annurev.genom.4.070802.110226
13. Novembre J, Stephens M. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*. 2008;40:646-649. doi:10.1038/ng.139
14. Marchani EE, Watkins WS, Bulayeva K, Harpending HC, Jorde LB. Culture creates genetic structure in the Caucasus: autosomal, mitochondrial, and Y-chromosomal variation in Daghestan. *BMC Genet*. 2008;9:47. doi:10.1186/1471-2156-9-47
15. Wightman DP, Jansen IE, Savage JE, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat Genet*. 2021;53:1276-1282. doi:10.1038/s41588-021-00921-z
16. Kunkle BW, Schmidt M, Klein H-U, et al. Novel Alzheimer disease risk loci and pathways in African American individuals using the African Genome Resources Panel: a meta-analysis. *JAMA Neurol*. 2021;78:102-113. doi:10.1001/jamaneurol.2020.3536
17. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet*. 2019;51:414-430. doi:10.1038/s41588-019-0358-2
18. Bellenguez C, Küçükali F, Jansen IE, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet*. 2022;54:412-436. doi:10.1038/s41588-022-01024-z
19. National Academies of Sciences and Medicine E. Using population descriptors in genetics and genomics research: a new framework for an evolving field. Washington, DC: National Academies Press, 2023.
20. OMB (U.S. Office of Management and Budget). Revisions to the standards for the classification of federeal data on race and ethnicity. 1997. https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf
21. McConkie-Rosell A, Spillmann RC, Schoch K, et al. Unraveling nonparticipation in genomic research: A complex interplay of barriers, facilitators, and sociocultural factors. *J Genet Couns*. 2023;32:993-1008. doi:10.1002/jgc4.1707
22. National Alzheimer's Project Act. Office of the Assistant Secretary for Planning and Evaluation (ASPE). n.d. Accessed July 31, 2023. https://aspe.hhs.gov/collaborations-committees-advisory-groups/napa
23. Mena PR, Kunkle BW, Faber KM, et al. The Alzheimer's Disease Sequencing Project Follow Up Study (ADSP-FUS): increasing ethnic diversity in Alzheimer's disease (AD) genetics research. *Alzheimers Dement*. 2022;18:e068083. doi:10.1002/alz.068083
24. Rajan KB, Weuve J, Barnes LL, Mcaninch EA, Wilson RS, Evans DA. Population estimate of people with clinical Alzheimer's disease and mild cognitive impairment in the United States (2020-2060). *Alzheimers Dement*. 2021;17:1966-1975. doi:10.1002/alz.12362
25. Matthews KA, Xu W, Gaglioti AH, et al. Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015-2060) in adults aged ≥65 years. *Alzheimers Dement*. 2019;15:17-24. doi:10.1016/j.jalz.2018.06.3063
26. Centers for Disease Control and Prevention NC for HStatistics. National Vital Statistics System, Mortality 1999-2020 on CDC WONDER Online Database, released in 2021. Data are from the Multiple Cause of Death Files, 1999-2020, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed January 10, 2023. http://wonder.cdc.gov/ucd-icd10.html
27. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet*. 2019;51:404-413. doi:10.1038/s41588-018-0311-9
28. Harold D, Abraham R, Hollingworth P, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet*. 2009;41:1088-1093. doi:10.1038/ng.440
29. Schwartzentruber J, Cooper S, Liu JZ, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat Genet*. 2021;53:392-402. doi:10.1038/s41588-020-00776-w
30. Jonsson T, Stefansson H, Steinberg S, et al. Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med*. 2012;368:107-116. doi:10.1056/NEJMoa1211103
31. Guerreiro R, Wojtas A, Bras J, et al. TREM2 variants in Alzheimer's disease. *N Engl J Med*. 2012;368:117-127. doi:10.1056/NEJMoa1211851
32. Chen H, Wu G, Jiang Y, et al. Analyzing 54,936 samples supports the association between CD2AP rs9349407 polymorphism and Alzheimer's disease susceptibility. *Mol Neurobiol*. 2015;52:1-7. doi:10.1007/s12035-014-8834-2
33. Miyashita A, Koike A, Jun G, et al. SORL1 is genetically associated with late-onset Alzheimer's disease in Japanese, Koreans and Caucasians. *PLoS One*. 2013;8:e58618. doi:10.1371/journal.pone.0058618
34. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526:68-74. doi:10.1038/nature15393
35. Tosto G, Fu H, Vardarajan BN, et al. F-box/LRR-repeat protein 7 is genetically associated with Alzheimer's disease. *Ann Clin Transl Neurol*. 2015;2:810-820. doi:10.1002/acn3.223
36. Cukier HN, Kunkle BW, Vardarajan BN, et al. ABCA7 frameshift deletion associated with Alzheimer disease in African Americans. *Neurol Genet*. 2016;2:e79. doi:10.1212/NXG.0000000000000079
37. Privé F, Aschard H, Carmi S, et al. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am J Hum Genet*. 2022;109:12-23. doi:10.1016/j.ajhg.2021.11.008
38. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51:584-591. doi:10.1038/s41588-019-0379-x

39. Baptiste, D, Caviness-Ashe, N, Josiah, N, et al. Henrietta Lacks and America's dark history of research involving African Americans. *Nursing Open*. 2022;9:2236–2238. Portico. doi:10.1002/nop2.1257
40. Jones BL, Vyhlidal CA, Bradley-Ewing A, Sherman A, Goggin K. If we would only ask: how Henrietta lacks continues to teach us about perceptions of research and genetic research among African Americans today. *J Racial Ethn Health Disparities*. 2017;4:735-745. doi:10.1007/s40615-016-0277-1
41. Claw KG, Anderson MZ, Begay RL, et al. A framework for enhancing ethical genomic research with Indigenous communities. *Nat Commun*. 2018;9:2957. doi:10.1038/s41467-018-05188-3
42. Ewing A, Thompson N, Ricks-Santi L. Strategies for enrollment of African Americans into cancer genetic studies. *J Cancer Educ*. 2015;30:108-115. doi:10.1007/s13187-014-0669-z
43. Dilworth-Anderson P, Hendrie HC, Manly JJ, Khachaturian AS, Fazio S. Diagnosis and assessment of Alzheimer's disease in diverse populations. *Alzheimers Dement*. 2008;4:305-309. doi:10.1016/j.jalz.2008.03.001
44. Gonzalez S, Strizich G, Isasi CR, et al. Consent for use of genetic data among US Hispanics/Latinos: results from the Hispanic Community Health Study/Study of Latinos. *Ethn Dis*. 2021;31:547. doi:10.18865/ed.31.4.547
45. Brown KE, Fohner AE, Woodahl EL. Beyond the individual: community-centric approaches to increase diversity in biomedical research. *Clin Pharmacol Ther*. 2023;113:509-517. doi:10.1002/cpt.2808
46. Boyer BB, Mohatt G V, Pasker RL, Drew EM, McGlone KK. Sharing results from complex disease genetics studies: a community based participatory research approach. *Int J Circumpolar Health*. 2007;66:19-30. doi:10.3402/ijch.v66i1.18221
47. Ochs-Balcom HM, Jandorf L, Wang Y, et al. "It takes a village": multilevel approaches to recruit African Americans and their families for genetic research. *J Community Genet*. 2015;6:39-45. doi:10.1007/s12687-014-0199-8
48. Wang L-S, Ho P-C, Tee BL, et al. The Asian Cohort for Alzheimer's Disease (ACAD) Pilot Study. *Alzheimers Dement*. 2022;18:e065599. doi:10.1002/alz.065599
49. Kang S, Gim J, Lee J, et al. Potential novel genes for late-onset Alzheimer's disease in East-Asian descent identified by APOE-stratified genome-wide association study. *J Alzheimers Dis*. 2021;82:1451-1460. doi:10.3233/JAD-210145
50. Yi D, Byun MS, Risacher SL, et al. The Korean brain aging study for the early diagnosis and prediction of Alzheimer's disease (KBASE): cognitive data harmonization. *Alzheimers Dement*. 2023;19:e064533. doi:10.4306/pi.2017.14.6.851
51. Lee J, Dey AB. Introduction to LASI-DAD: the longitudinal aging study in India-diagnostic assessment of dementia. *J Am Geriatr Soc*. 2020;68:S3. doi:10.1111/jgs.16740
52. Ruiz NG, Noe-Bustamante L, Shah S. Appendix: demographic profile of Asian American adults. Pew Research Center. 2023. https://www.pewresearch.org/race-ethnicity/2023/05/08/asian-american-identity-appendix-demographic-profile-of-asian-american-adults/#:~:text=About%2017.8%20million%20Asian%20adults,adults%20lived%20in%20the%20country
53. Riehm KE, Keyes KM, Susser ES. Social determinants of health and selection bias in genome-wide association studies. *World Psychiatry*. 2023;22:160. doi: 10.1002/wps.21047
54. Shriner D. Overview of admixture mapping. *Curr Protoc Hum Genet*. 2013. Chapter 1. doi:10.1002/0471142905.hg0123s76
55. Horimoto ARVR, Xue D, Thornton TA, Blue EE. Admixture mapping reveals the association between Native American ancestry at 3q13.11 and reduced risk of Alzheimer's disease in Caribbean Hispanics. *Alzheimers Res Ther*. 2021;13:122. doi:10.1186/s13195-021-00866-9
56. Horimoto ARVR, Boyken LA, Blue EE, et al. Admixture mapping implicates 13q33. 3 as ancestry-of-origin locus for Alzheimer disease in Hispanic and Latino populations. *HGG Adv*. 2023;4:100207. doi: 10.1016/j.xhgg.2023.100207
57. Kizil C, Sariya S, Kim YA, et al. Admixture mapping of Alzheimer's disease in Caribbean Hispanics identifies a new locus on 22q13. 1. *Mol Psychiatry*. 2022;27:2813-2820. doi: 10.1038/s41380-022-01526-6
58. Atkinson EG, Maihofer AX, Kanai M, et al. Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat Genet*. 2021;53:195-204. doi:10.1038/s41588-020-00766-y
59. Rajabli F, Benchek P, Tosto G, et al. Multi-ancestry genome-wide meta-analysis of 56,241 individuals identifies LRRC4C, LHX5-AS1 and nominates ancestry-specific loci PTPRK, GRB14, and KIAA0825 as novel risk loci for Alzheimer disease: the Alzheimer Disease Genetics Consortium. *MedRxiv*. 2023:2023-2027.
60. Clark PC, Kutner NG, Goldstein FC, et al. Impediments to timely diagnosis of Alzheimer's disease in African Americans. *J Am Geriatr Soc*. 2005;53:2012-2017. doi:10.1111/j.1532-5415.2005.53569.x
61. Griffin-Pierce T, Silverberg N, Connor D, et al. Challenges to the recognition and assessment of Alzheimer's disease in American Indians of the southwestern United States. *Alzheimers Dement*. 2008;4:291-299. doi:10.1016/j.jalz.2007.10.012
62. Barnes LL. Alzheimer disease in African American individuals: increased incidence or not enough data? *Nat Rev Neurol*. 2022;18:56-62. doi:10.1038/s41582-021-00589-3
63. Ski CF, King-Shier KM, Thompson DR. Gender, socioeconomic and ethnic/racial disparities in cardiovascular disease: a time for change. *Int J Cardiol*. 2014;170:255-257. doi:10.1016/j.ijcard.2013.10.082
64. O'Keefe EB, Meltzer JP, Bethea TN. Health disparities and cancer: racial disparities in cancer mortality in the United States, 2000-2010. *Front Public Health*. 2015;3:51. doi:10.3389/fpubh.2015.00051
65. Mayeda ER, Glymour MM, Quesenberry CP, Johnson JK, Pérez-Stable EJ, Whitmer RA. Survival after dementia diagnosis in five racial/ethnic groups. *Alzheimers Dement*. 2017;13:761-769. doi:10.1016/j.jalz.2016.12.008

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Xue D, Blue EE, Conomos MP, Fohner AE. The power of representation: Statistical analysis of diversity in US Alzheimer's disease genetics data. *Alzheimer's Dement*. 2024;10:e12462. https://doi.org/10.1002/trc2.12462