nature genetics

Article

https://doi.org/10.1038/s41588-024-01963-9

Pervasive biases in proxy genome-wide association studies based on parental history of Alzheimer's disease

Received: 26 October 2023

Accepted: 27 September 2024

Published online: 4 November 2024

Check for updates

Yuchang Wu[1,5], Zhongxuan Sun[1,5], Qinwen Zheng[1], Jiacheng Miao[1], Stephen Dorn[1], Shubhabrata Mukherjee[2], Jason M. Fletcher[3,4] & Qiongshi Lu[1]

Almost every recent Alzheimer's disease (AD) genome-wide association study (GWAS) has performed meta-analysis to combine studies with clinical diagnosis of AD with studies that use proxy phenotypes based on parental disease history. Here, we report major limitations in current GWAS-by-proxy (GWAX) practices due to uncorrected survival bias and nonrandom participation in parental illness surveys, which cause substantial discrepancies between AD GWAS and GWAX results. We demonstrate that the current AD GWAX provide highly misleading genetic correlations between AD risk and higher education, which subsequently affects a variety of genetic epidemiological applications involving AD and cognition. Our study sheds light on potential issues in the design and analysis of middle-aged biobank cohorts and underscores the need for caution when interpreting genetic association results based on proxy-reported parental disease history.

Genome-wide association studies (GWAS) have greatly advanced our understanding of the genetic underpinning of complex diseases, revealing numerous genotype–phenotype associations[1]. This progress is largely driven by population biobanks, such as the UK Biobank (UKB)[2], which provide extensive genotype and phenotype data on large samples. However, a persistent challenge in biobank-based GWAS applications is that study participants are typically middle-aged without late-onset disease diagnoses. To address this limitation, Liu et al.[3] introduced GWAS-by-proxy (GWAX) based on a simple idea—although biobank participants may not have their own diagnoses on late-life disease outcomes, they report their parents' diagnoses through family health history surveys and they also (indirectly) provide parental genetic data as their biological children. Liu et al. demonstrated the efficacy of GWAX by replicating risk loci from case–control GWAS for several diseases, including Alzheimer's disease (AD)[3]. Since then, GWAX has quickly gained popularity in complex disease genetic research, especially for neurodegenerative diseases. GWAX has become so popular in AD genetic studies that every recent AD GWAS performed meta-analysis to combine associations from clinically diagnosed AD cases versus controls[4] with GWAX associations to boost statistical power[5–10]. Notably, the largest AD GWAS to date[9] did not share separate association results for GWAS and GWAX in their study. Instead, only the meta-analyzed association results were made available to the research community.

However, methodological issues in GWAX and the quality of its association results remain underexplored. Liu et al.[3] provided evidence that top significant loci yielded similar results in GWAS and GWAX. Since then, critiques of GWAX have mostly focused on the imprecision of survey data (that is, measurement error in parental health history) and the implications in genetic applications (for example, heritability estimation)[11,12]. Few studies have investigated potential systematic biases and methodological limitations in GWAX, particularly regarding the infinitesimal biases that do not appear substantial when focusing on top GWAS loci with large effects but that could

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI, USA. [2]Department of Medicine, University of Washington, Seattle, WA, USA. [3]Department of Population Health Sciences, University of Wisconsin–Madison, Madison, WI, USA. [4]La Follette School of Public Affairs, University of Wisconsin–Madison, Madison, WI, USA. [5]These authors contributed equally: Yuchang Wu, Zhongxuan Sun.
e-mail: qlu@biostat.wisc.edu

https://doi.org/10.1038/s41588-024-01963-9

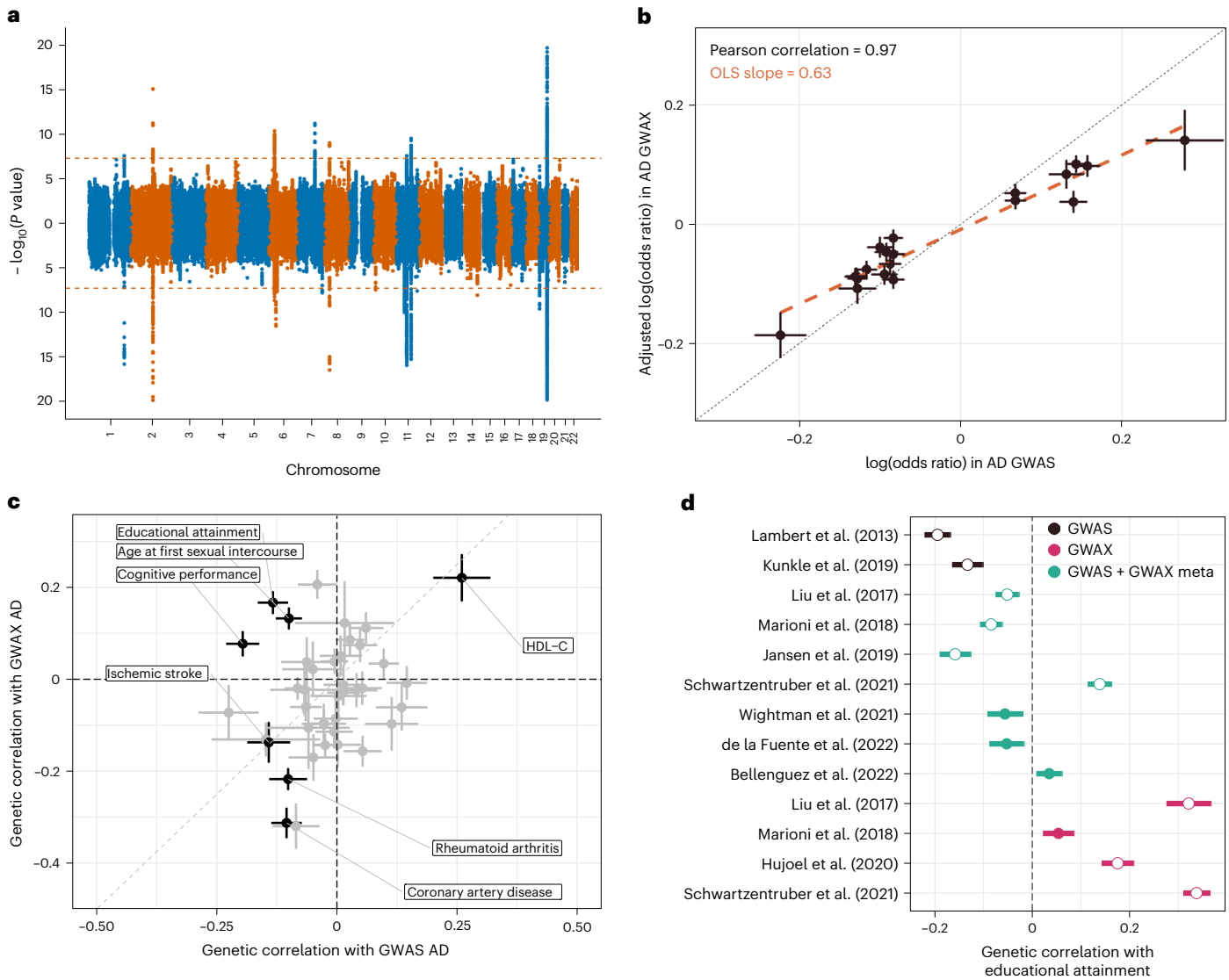

Fig. 1 | Comparing top association findings and genetic correlation results of AD GWAS and GWAX. a, Manhattan plot for AD GWAX (top) and GWAS (Kunkle et al.[13]; bottom). The *y* axis represents the negative base-10 logarithm of the two-sided raw *P* values from logistic regression score tests and is capped at 20 for better visualization. Horizontal lines mark the genome-wide significance threshold ($P = 5.0 \times 10^{-8}$). **b**, GWAS and GWAX effect size estimates for 20 genome-wide-significant SNPs identified by Kunkle et al.[13]. The GWAX effect sizes were adjusted using the genetic relatedness of 0.5 between parents and children. The *APOE* locus was excluded due to its extreme effect size. Orange dotted line represents the fitted line from ordinary least squares (OLS) regression. **c**, Genetic correlations of the AD GWAX and GWAS with 40 complex traits. Traits with significant correlations (FDR < 0.05) with both AD GWAS and GWAX are highlighted and labeled. HDL-C, high-density lipoprotein cholesterol. **d**, Genetic correlation of AD and EA based on ten AD genetic studies published between 2013 and 2022. In **b**–**d**, dots and intervals indicate point estimates and ±1 s.e. for the estimates, respectively. Significant results at an FDR cutoff of 0.05 are highlighted with white circles in **d**.

severely bias applications that involve genome-wide association estimates, such as genetic correlation estimation and polygenic risk score (PRS).

Here, we report evidence of widespread discrepancies between GWAX based on family history and case–control GWAS for AD[13], revealing pervasive biases in current GWAX approaches. We implemented GSUB, a GWAS-by-subtraction strategy[14], to quantify the biases originating from different sources, and revealing that AD GWAX suffers from survival bias from differential parental lifespans and participation and reporting biases in parental health history surveys. We demonstrate that almost all existing GWAX approaches produce counterintuitive positive associations between higher cognition/education and dementia risk. We show that several genetic epidemiological applications involving AD and cognition yield mixed findings due to this issue in existing studies. We also use a variety of methods to reduce these biases and benchmark their performances. Our findings emphasize the urgent need for caution when interpreting GWAX results and provide guidance for future study designs involving proxy phenotypes derived from family health history.

Results

GWAX replicates AD loci but diverges in genetic correlations

To assess the validity of GWAX for AD, we aimed to replicate the genome-wide-significant loci ($P \leq 5 \times 10^{-8}$) identified in a recent AD case–control study[13]. We performed GWAX using UKB participants of European descent who reported parental history of AD or dementia (n = 47,993 proxy cases and 315,096 controls; Methods). AD GWAX produced similar results compared to GWAS (Fig. 1a,b and Supplementary Table 1). Consistent with previous findings[3], GWAS and GWAX effect estimates were highly correlated (cor. = 0.97 with *APOE* excluded),

https://doi.org/10.1038/s41588-024-01963-9

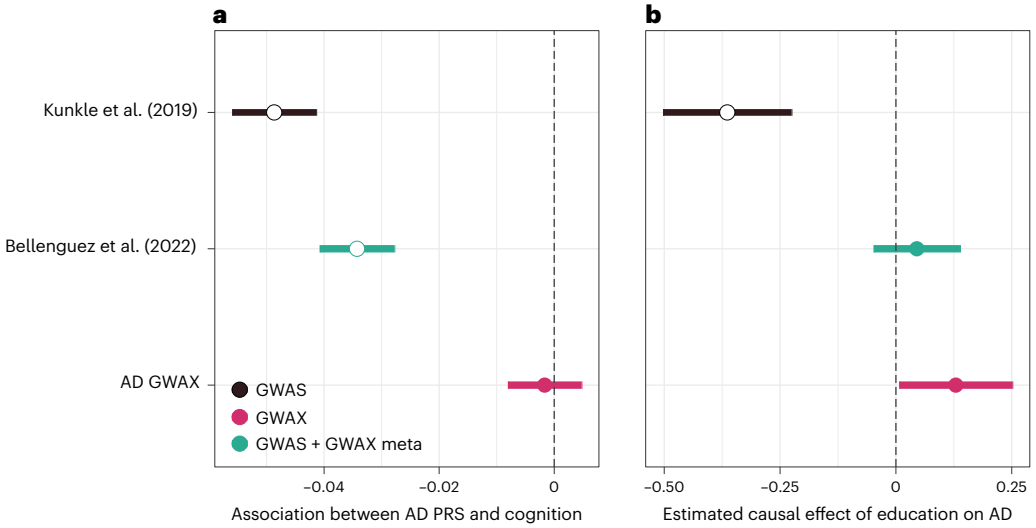

Fig. 2 | AD GWAX biases risk prediction and causal inference. a, Association of AD PRS and late-life cognition in the HRS cohort. PRS were computed from genome-wide association results using the PRS-CS approach. **b**, Causal effect of EA on AD risk estimated from Mendelian randomization. For both panels, the dots and intervals indicate point estimates and ±1 s.e. for the estimates, respectively. Significant results at an FDR cutoff of 0.05 are shown as white circles. Data for this plot are in Supplementary Tables 6 and 7.

but GWAX showed substantial attenuation in effect sizes (regression slope = 0.63). This attenuation was not explained by measurement errors (Supplementary Fig. 1) but may be explained by winner's curse: the regression slope increased to 1.15 (standard error (s.e.) = 0.20) after correction[15] (Methods). Similar results were found using the top single-nucleotide polymorphisms (SNPs) from GWAX (Supplementary Table 2 and Supplementary Fig. 1).

Discrepancies between GWAX and GWAS became evident in analyses leveraging genome-wide data that included SNPs not reaching Bonferroni-corrected statistical significance. We estimated the genetic correlations of GWAS and GWAX AD with 40 complex traits (Supplementary Tables 3 and 4). AD GWAS and GWAX were significantly correlated (cor. = 0.63, $P = 3.9 \times 10^{-31}$), but they showed divergent correlations with multiple traits (Supplementary Fig. 2). For example, total cholesterol and hippocampal volume showed significant genetic correlations with AD GWAS (cor. = 0.13 and −0.23; $P = 0.01$ and 3.1×10^{-4}) but not with AD GWAX (cor. = −0.061 and −0.073; $P = 0.23$ and 0.23). Attention-deficit/hyperactivity disorder (ADHD) and coronary artery disease showed substantially stronger correlations with lower AD risk in GWAX (cor. = −0.16 and −0.31; $P = 2.9 \times 10^{-6}$ and 3.4×10^{-21}) than in GWAS (cor. = 0.05 and −0.1; $P = 0.18$ and 9.5×10^{-4}). Excluding the *APOE* region (chr19:45116911–46318605; GRCh37) did not substantially affect these results (Supplementary Fig. 3).

Only seven traits had significant correlations with both AD GWAS and GWAX under a false discovery rate (FDR) cutoff of 0.05, with three correlations flipping direction (Fig. 1c). In particular, educational attainment (EA), a well-documented negative correlate of AD risk[16,17], showed an expected negative genetic correlation with AD GWAS (cor. = −0.13, $P = 2.4 \times 10^{-5}$) but a significant positive correlation with AD GWAX (cor. = 0.17, $P = 1.7 \times 10^{-11}$). To determine whether this was a consistent finding across studies, we summarized the genetic correlations between EA and AD in ten AD studies published from 2013 to 2022 (refs. 3,5–9,13,18–20) (Supplementary Table 5). All the studies showed a consistent pattern: higher education was correlated with lower AD risk in case–control studies but was correlated with higher risk in family history-based studies, with GWAS–GWAX meta-analysis results falling in between (Fig. 1d). The largest AD genetic association study to date[9], a meta-analysis of GWAS and GWAX, showed a positive genetic correlation with EA, but this did not reach statistical significance (cor. = 0.03, $P = 0.18$).

GWAX biases risk prediction and causal inference

Given the concerningly divergent AD–EA genetic correlations from GWAS and GWAX, we investigated two genetic epidemiological applications involving AD and cognition. First, we quantified the predictive performance of AD PRS on late-life cognition in the Health and Retirement Study (HRS). We calculated three PRS from AD GWAS[13], GWAX and a GWAS–GWAX meta-analysis[9] and associated these scores with the global cognition composite score while controlling for key covariates ($n = 12{,}018$; Methods). The GWAS-based PRS exhibited a strong association with lower cognition (effect = −0.05, $P = 2.5 \times 10^{-11}$) while the GWAX-based PRS did not show significant association (effect = −0.0017, $P = 0.80$; Fig. 2a). PRS based on the Bellenguez et al.[9] meta-analysis was associated with lower cognition but showed an attenuated effect (effect = −0.03, $P = 1.2 \times 10^{-7}$), despite the substantially larger sample size. Similar results were found after removing *APOE* from all PRS (Supplementary Fig. 4). GWAS-based and GWAX-based PRS using only genome-wide-significant variants showed similar performance (Supplementary Fig. 4), suggesting that biases in PRS were mostly driven by SNPs not reaching genome-wide significance.

Education has been hypothesized to have a causal protective effect against AD. Many studies have investigated this hypothesis with mixed results[16,17,21,22]. Using Mendelian randomization, we estimated the causal effect of EA on AD risk (Methods). Again, we observed inconsistent results between AD GWAS and GWAX (Fig. 2b). We identified a significant protective effect of EA on AD risk using AD GWAS (effect = −0.36, $P = 8.6 \times 10^{-3}$). In contrast, when GWAX was the outcome study, EA was estimated to increase AD risk, although the effect was not statistically significant. A slightly positive but nonsignificant causal effect of EA on AD risk was also identified using the meta-analysis from Bellenguez et al.[9]. Sensitivity analyses confirmed that these results were robust to pleiotropy and effect heterogeneity (Methods and Supplementary Table 7). The discrepancies between AD GWAS and GWAX in these analyses underscore the need to reevaluate GWAX applications in AD genetic studies.

GWAS-by-subtraction identifies potential biases in AD GWAX

Next, we applied a GWAS-by-subtraction approach to separate biases from AD genetic associations in GWAX. This approach assumes that GWAX associations can be explained by AD signals (that is, AD factor F_{AD}) and biases (that is, non-AD factor F_{non}). It quantifies the genetic basis of

https://doi.org/10.1038/s41588-024-01963-9

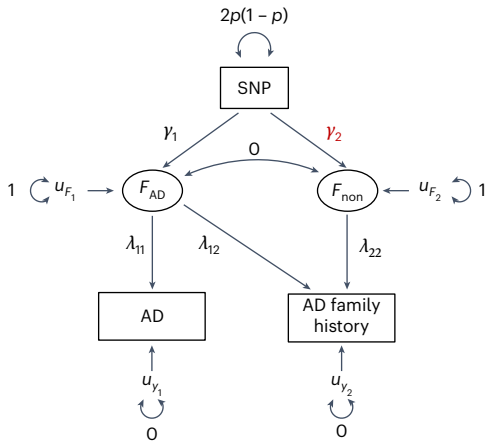

Fig. 3 | Schematic diagram for GWAS-by-subtraction. The main goal is to estimate genetic associations γ_2 with the nondisease factor F_{non} underlying parental disease history. $\gamma_{1,2}$ and $\lambda_{11,12,22}$ are the parameters that need to be estimated (Methods).

the non-AD component by regressing out AD case–control associations from GWAX results (Methods and Fig. 3). GWAS-by-subtraction[14,23,24] has had several important applications in the literature and is implemented under GenomicSEM[25]. We introduce an alternative strategy for GWAS-by-subtraction based on our previous work[26], due to computational singularity issues caused by the high genetic correlation between AD GWAS and GWAX. This approach produces closed-form estimates for the main parameters of interest, that is, SNP effects on the non-AD factor F_{non} (Fig. 3). We implemented this approach in a software package named GSUB (Methods). Compared to GenomicSEM, GSUB produces consistent effect estimates with comparable statistical power without convergence issues and is computationally much faster (Supplementary Fig. 5 and Supplementary Table 8).

To elucidate the mechanisms behind the non-AD (that is, bias) genetic component underlying AD GWAX, we computed the genetic correlations of the non-AD component with 50 complex traits (Supplementary Tables 9 and 10). These included 40 complex traits from previous analyses. Due to EA's divergent genetic correlations with the AD GWAS and GWAX (Fig. 1c,d), we also included three additional GWAS targeting different aspects of EA and cognition: direct and indirect (that is, parental) genetic effects on EA estimated from family-based GWAS[26] and the noncognitive component for EA[14]. Further, to compare with non-AD dementia, we included GWAS for Parkinson's disease[27], amyotrophic lateral sclerosis[28], frontotemporal dementia[29] and Lewy body dementia[30]. Finally, to investigate the effect of nonrandom participation, we performed GWAS on 'Do not know parental illness' in UKB (Supplementary Fig. 6 and Supplementary Table 11), as well as family medical history awareness and participation in the family health history survey (Supplementary Figs. 7–10 and Supplementary Tables 12 and 13) using data from the All of Us Research Program (AllofUs) (Methods). This increased the total number of traits to 50.

Figure 4 shows significant genetic correlations with the non-AD component underlying GWAX; 16 traits reached statistical significance at FDR < 0.05. The full results for the 50 traits are shown in Supplementary Fig. 11. The non-AD component exhibited substantial correlations with higher EA (cor. = 0.26, $P = 5.2 \times 10^{-11}$), indirect (parental) effect on EA (cor. = 0.53, $P = 4.5 \times 10^{-4}$), cognition (cor. = 0.19, $P = 1.4 \times 10^{-4}$) and the noncognitive component[14] for EA (cor. = 0.23, $P = 1.4 \times 10^{-6}$). We also observed negative genetic correlations with health outcomes such as major depressive disorder (cor. = −0.11, $P = 0.012$), schizophrenia (cor. = −0.13, $P = 8.3 \times 10^{-3}$), coronary artery disease (cor. = −0.14, $P = 2.8 \times 10^{-3}$), ADHD (cor. = −0.17, $P = 0.012$), epilepsy (cor. = −0.20, $P = 7.1 \times 10^{-4}$) and heart failure (cor. = −0.24, $P = 4.1 \times 10^{-4}$), suggesting potential survival bias in AD GWAX. That is, parents who have an AD diagnosis would have to have lived long enough to receive the diagnosis, thus having lower genetic risks for other health issues due to competing risks. Meanwhile, younger parents who have not reached the age of dementia onset will not have lower genetic risks for other outcomes. Therefore, the genetic footprints for many health outcomes could partially explain the genetic differences between proxy cases and controls. Indeed, we observed distinct age distributions for GWAX cases and controls (Supplementary Fig. 12). Compared to proxy cases, participants who did not report parental AD history, along with their parents, were younger.

Because the survey question about parental health history in UKB, that is, "Has/did your father (mother) ever suffer from Alzheimer's disease/dementia?", lacks clear differentiation between AD and other dementias, we examined whether genetic associations for non-AD dementia could explain the biases in AD GWAX. All four non-AD dementias included in our analysis showed null results with genetic correlations close to zero (Supplementary Table 9 and Supplementary Fig. 11), providing very limited evidence for this hypothesis.

A recent study[31] demonstrated a genetic basis for nonrandom survey responses in UKB. We investigated whether participation in the family health history survey and systematic misreporting of parental disease status could explain the biases in AD GWAX. We found significant genetic correlations between the non-AD component and family health history survey participation (cor. = 0.44, $P = 1.1 \times 10^{-9}$) and (not) knowing parental illnesses (cor. = −0.18, $P = 4.6 \times 10^{-3}$).

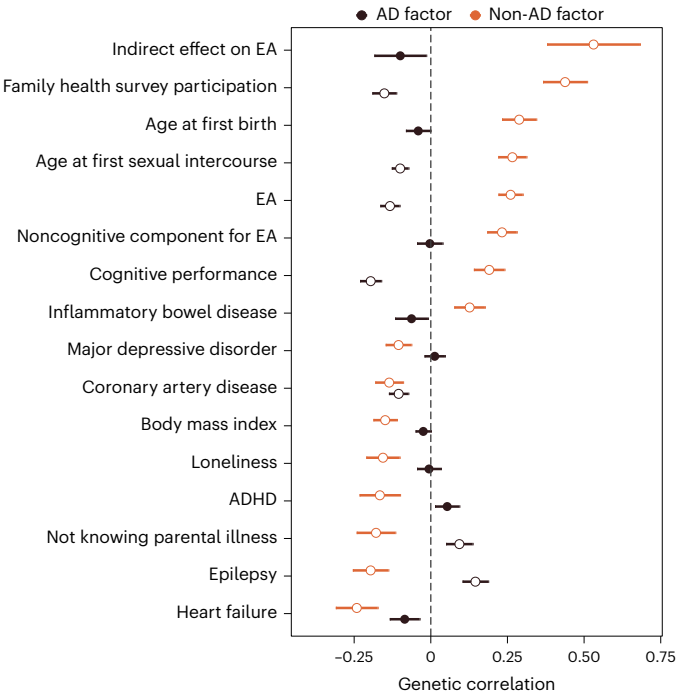

Fig. 4 | Genetic correlation of the AD and non-AD factors in GWAX with other complex traits. The GWAS used for the AD factor was the case–control GWAS by Kunkle et al.[13] (Methods), and genetic associations with the non-AD component were obtained using GWAS-by-subtraction. The plot shows 16 traits having significant correlations with the non-AD factor. Dots and intervals indicate point estimates and ±1 s.e. for the estimates, respectively. Significant correlations with FDR < 0.05 are shown as white circles. The full genetic correlation results are reported in Supplementary Tables 3, 9 and 10.

https://doi.org/10.1038/s41588-024-01963-9

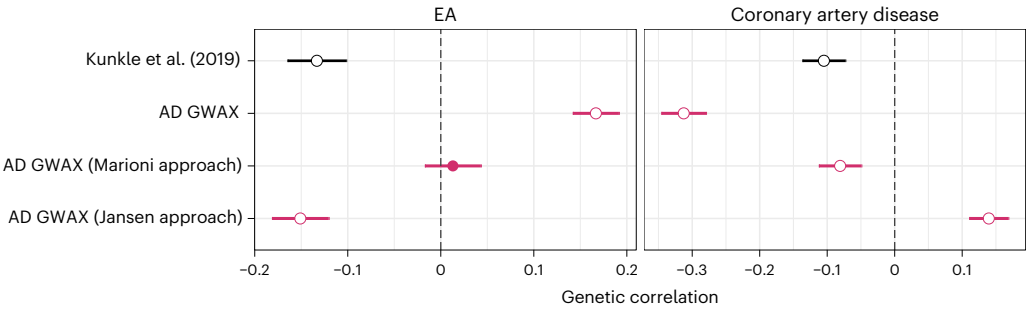

Fig. 5 | Genetic correlation of AD GWAS and GWAX with EA and coronary artery disease. We included two approaches to correct for survival bias. Following Marioni et al.[5], we required participants' parental age (either current age or age at death) to be older than the AD onset age of 65 and included parental age in the covariates (Methods). Following Jansen et al.[6], we ran a GWAS on a continuous phenotype that was constructed based on parental AD status, parental age and AD prevalence to quantify the disease load. Dots and intervals indicate the point estimates and ±1 s.e. for the estimates, respectively. Significant results at an FDR cutoff of 0.05 are shown as white circles. Data for this plot are in Supplementary Tables 14 and 15.

Reducing biases in GWAX

Having identified potential sources of bias in AD GWAX, we explored various methods to reduce these biases. To reduce survival bias, we applied two approaches to control for parental age and vital status in the regression (Supplementary Table 14). Following Marioni et al.[5], we excluded participants with parents younger than 65 years and added parental age as a covariate in GWAX; following Jansen et al.[6], we constructed a continuous GWAX phenotype using parental AD status, parental age and AD prevalence (Methods). Using AD–EA genetic correlation as a benchmark, both approaches reduced bias (Fig. 5). In the Marioni approach, AD showed a null genetic correlation with EA, whereas the Jansen approach flipped the AD–EA genetic correlation from 0.17 to −0.15, closely matching the correlation in the AD case–control GWAS (cor. = −0.13). We also examined the genetic correlation with coronary artery disease as a benchmark for survival bias (Fig. 5). The Marioni approach substantially reduced the genetic correlation (cor. = −0.081, $P = 0.01$), yielding a similar result as the AD case–control GWAS. In the Jansen approach, AD had a significant positive genetic correlation with coronary artery disease (cor. = 0.14, $P = 1.5 \times 10^{-6}$).

To reduce participation bias, we followed the approach of Schoeler et al.[32] and conducted two sets of weighted GWAS on parental AD status. First, we trained a LASSO regression model on whether a survey participant reported parental illnesses using a random subset of UKB samples and performed weighted GWAS on the remaining samples (Methods). However, this approach did not improve the genetic correlation estimates with EA (cor. = 0.19, $P = 4.3 \times 10^{-5}$) or coronary artery disease (cor. = −0.41, $P = 2.1 \times 10^{-7}$; Supplementary Tables 14 and 15). Switching to sample weights contrasting UKB participants and the general UK population (Methods) also did not improve the genetic correlation results for EA (cor. = 0.18, $P = 1.2 \times 10^{-5}$) or coronary artery disease (cor. = −0.37, $P = 1.2 \times 10^{-8}$). We also explored using GWAS-by-subtraction to adjust for participation bias by regressing out the participation GWAS performed in AllofUs from AD GWAX (Methods). The residual GWAS showed reduced genetic correlations with EA (cor. = 0.11, $P = 1.1 \times 10^{-4}$) and coronary artery disease (cor. = −0.18, $P = 1.7 \times 10^{-9}$), but both correlations remained statistically significant (Supplementary Tables 14 and 15).

To address the bias due to systematic over-reporting or under-reporting in the parental health survey, we explored two strategies. First, in our default GWAX implementation, we excluded people who reported 'do not know' about parental health, thus already controlling for family health awareness to some extent. To investigate whether this is a reasonable strategy, we implemented another GWAX including people who did not know their parents' health as controls. As expected, this increased genetic correlations between AD and higher EA (cor. = 0.21, $P = 7.7 \times 10^{-19}$; Supplementary Table 14) and lower risk for coronary artery disease (cor. = −0.33, $P = 2.5 \times 10^{-26}$; Supplementary Table 15). We also applied GWAS-by-subtraction to regress out the 'do not know parental illness' genetic component from AD GWAX (Methods and Supplementary Fig. 13). The residual GWAS showed substantially reduced yet still significant correlations with EA (cor. = 0.07, $P = 0.027$) and coronary artery disease (cor. = −0.26, $P = 1.3 \times 10^{-13}$).

Finally, we note that, although both the Marioni and Jansen approaches were primarily designed for reducing survival bias alone, they also removed some reporting bias. After correction, AD GWAX had null genetic correlations with 'do not know parental illness' (cor. = −0.03 and 0.087, $P = 0.54$ and 0.06 for the Marioni and Jansen approaches, respectively; Supplementary Tables 16 and 17).

Meta-analysis of GWAS and GWAX associations

AD GWAX is often meta-analyzed with clinically diagnosed case–control GWAS to boost statistical power. We investigated whether accounting for heterogeneity when meta-analyzing GWAS and GWAX could reduce biases in the combined association results. We explored two approaches: METAL[33] is a common approach for meta-analysis and GenomicSEM was recently proposed as an alternative strategy that can account for measurement error and phenotype heterogeneity in GWAX–GWAS meta-analyses[12,20]. Figure 6 illustrates the genetic correlations of the meta-analyzed outcomes based on the two meta-analytic approaches with EA and coronary artery disease. Compared to the results in Fig. 5, meta-analyzing GWAX with GWAS produced genetic correlations somewhere between those given by GWAX and GWAS. Meta-analysis alone cannot sufficiently remove all the bias. The two meta-analytic methods produced mostly comparable results, highlighting the importance of reducing biases in GWAX analysis instead of relying solely on post hoc bias reduction during meta-analysis.

Discussion

In recent years, GWAX has emerged as a crucial study design for complex trait genetics in general and AD genetic research in particular, gaining popularity due to its ability to leverage middle-aged populations to study late-onset outcomes. The validity of GWAX was supported by two types of evidence in the literature: similar effect size estimates for top SNPs and high genetic correlation between GWAS and GWAX. Some concerns have been raised concerning the GWAX design, mostly focusing on measurement errors in family health history surveys. Escott-Price and Hardy[11] argued that parental AD cases inferred from vaguely defined surveys may encompass both AD and non-AD dementia cases, each with distinct genetic underpinnings, which would attenuate genuine genetic associations for AD. Grotzinger et al.[12] demonstrated that naively combining GWAS and GWAX without accounting for heterogeneity among the associations leads to substantial downward bias in heritability estimation. Despite these critiques, GWAX remains essential to AD studies[5–10], raising

https://doi.org/10.1038/s41588-024-01963-9

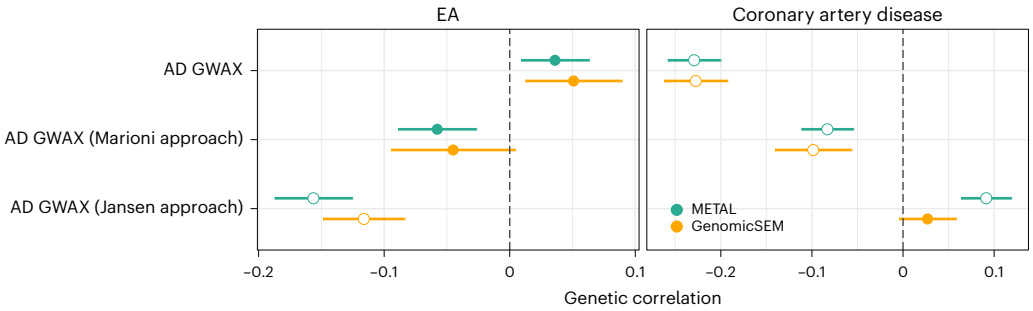

Fig. 6 | Genetic correlation of meta-analyzed AD with EA and coronary artery disease. Meta-analysis results based on the GWAS from Kunkle et al.[13] and three sets of AD GWAX are shown. Results based on two meta-analysis approaches (METAL[33] and GenomicSEM[25]) are also compared. We used METAL to combine associations from GWAX based on parental AD history with GWAS associations. Because GenomicSEM requires at least three studies as input, we meta-analyzed GWAS, paternal GWAX and maternal GWAX. Dots and intervals indicate point estimates and ±1 s.e. for the estimates, respectively. Significant correlations with FDR < 0.05 are shown as white circles. Data for this plot are in Supplementary Tables 14 and 15.

concerns about the quality of reported associations and the prospect of follow-up studies based on GWAS findings.

Our study revealed pervasive biases in AD GWAX. In particular, GWAX yielded an unexpected positive genetic correlation with EA, and such biases are present in almost all published AD GWAS that included proxy AD cases. The implications of this issue are twofold. First, the biases identified in our analyses are not speculation of some negligible issue in empirical applications. We demonstrated substantial divergence of AD GWAS and GWAX due to these biases. Second, an important social factor at the center of many of these biases is education, which is known to be associated with longevity, the parent–child relationship and general health awareness[34]. Because cognition is such a crucial marker for AD and is commonly used in dementia research, biases caused by education and cognition become particularly important in AD genetic research and may give misleading results if not handled properly, complicating diagnosis, treatment and the design and testing of new drugs. We investigated two popular genetic epidemiological analyses involving AD: predicting late-life cognition using AD PRS and estimating the causal (protective) effect of education on AD using Mendelian randomization. Both analyses were substantially influenced by biases in GWAX. This is alarming because of the considerable interest in identifying modifiable factors to reduce AD risk. Numerous studies have used Mendelian randomization to investigate the hypothesized protective role of education against AD and have reported mixed results. Our findings have highlighted GWAX as a source of heterogeneity in causal effect estimation. Indeed, while the studies[16,17] using AD GWAS as the outcome found a protective effect of education on AD, those[21,22] using AD GWAX reported counterintuitive results that education may elevate AD risk. We also compared the genetic correlations of EA with GWAS and GWAX for ten other diseases and found ubiquitously divergent results (Supplementary Fig. 14 and Supplementary Tables 19 and 20). This suggests that the issues we reported in this study may be present across a range of disease outcomes and need to be carefully examined in future studies.

Despite the strong evidence for bias in AD GWAX, the source of such bias was not clearly understood. We hypothesized that three mechanisms contribute to the bias. First, only people with parents who lived long enough can report parental AD diagnosis. Without adjusting for survival bias, we expect to see spurious negative genetic correlations between AD GWAX and other health outcomes. That is, genetic variants that are protective for other diseases will appear to increase the risk of AD because they are associated with longevity. Second, people who are more aware of their parents' health are more likely to report parental AD diagnosis. This could be affected by people's general awareness of health, but it may also be explained by people's relationship with their parents, whether they grew up in single-parent families, parents' socioeconomic status and other complex socioenvironmental factors. Third, parental AD cases reported in the UKB parental health survey may include non-AD dementia cases. Therefore, we expect genetic associations with other types of dementia to explain some of the differences between AD GWAS and GWAX. Using an innovative GWAS-by-subtraction strategy[14] and our closed-form implementation with superior robustness and computational efficiency, we quantified the genetic effects underlying AD GWAX that are not explained by genuine AD associations. We found substantial evidence for survival bias, supported by negative genetic correlations of the non-AD (bias) component with many health outcomes. We also found genetic correlations with survey participation and awareness of parental health history, which suggest nonrandom participation and reporting in the UKB survey as possible sources of bias. We did not find evidence for associations with other types of dementia in AD GWAX, although this is possibly explained by the lower statistical power in current non-AD dementia studies.

We investigated several approaches to reduce biases in AD GWAX. Controlling for parental age and vital status could effectively reduce survival bias. In particular, the approach that creates a continuous disease risk phenotype based on parental age[6] produced a genetic correlation with education comparable to that in AD case–control GWAS. However, one potential limitation of this approach is that it does not produce SNP effect sizes on a scale similar to that of case–control studies, which creates challenges in interpretation and some applications requiring effect sizes. While weighted least squares is a common approach to account for nonrandom participation[32], it did not give promising results in our analysis. Excluding individuals who do not know about their parents' health from the analysis and residualizing GWAX on genetic associations with parental health awareness both reduced the spurious AD–EA genetic correlation. We note that the approaches designed to remove survival bias also reduced some participation and reporting bias, suggesting entangled mechanisms behind these possibly over-simplified labels for different sources of bias. This provides a potential one-stop solution to multiple sources of bias, but its effectiveness remains to be investigated in the future. We also note that these biases could not be mitigated by a simple meta-analysis with AD case–control GWAS, further highlighting the importance of improving GWAX quality. Recently, several studies introduced methods to jointly model study participants' disease status and family history in GWAS analysis. For example, LT-FH[19] first calculates a continuous posterior disease liability conditioning on case–control status and family history and then conducts a GWAS on this liability score. We explored its performance on AD (Fig. 1d) and found that it produced a significant and positive genetic correlation with EA. Therefore, joint modeling of the self-reported disease status of individuals and their parents without consideration of potential sources of bias in

family history variables will most likely fail to produce valid association results. LT-FH++ (ref. 35) is an extension of LT-FH that accounts for age of disease onset when computing the posterior disease liability. The inclusion of age-of-onset information may help address survival bias. However, while UKB provides study participants' age-of-onset information, this information is not available for their parents. Without age of onset, LT-FH++ degenerates to LT-FH. It remains an open question whether LT-FH++ could produce unbiased results if the parental age of disease onset were to become available in the future. Finally, besides the issues we have detailed in this study, many association mapping approaches being used in GWAX studies appear to be poorly justified statistically. For example, AD GWAX sometimes combine clinically diagnosed cases and proxy cases in logistic regression without properly scaling the SNP effect size[7,9]. Some studies use both sibling and parental proxy cases[3,7,9], which could introduce additional survival bias and other complications. Some other studies meta-analyze GWAX associations based on maternal and paternal AD histories without accounting for sample overlap[5]. There is an urgent need to improve the general statistical methodology for handling family history outcomes in genetic association studies.

Our study had several limitations. First, we treated the AD case–control GWAS as the gold standard, but it remains plausible that some issues could affect the analysis based on AD clinical diagnosis. For example, the significant genetic correlation between AD GWAS by Kunkle et al.[13] and lower risk for coronary artery disease (Fig. 5) suggests an uncorrected survival bias in AD GWAS. In fact, GWAX following the Jansen approach showed a positive genetic correlation with coronary artery disease. It is unclear whether this is caused by limitations in the bias-removal approach or by correctly recovering the shared genetics between AD and cardiovascular disease risk[36,37]. We compared the GWAS by Kunkle et al.[13] with a GWAS of late-onset AD from FinnGen (Supplementary Table 21), which is a population cohort known to be less affected by sampling biases due to its recruitment strategies[38]. However, AD status in FinnGen was derived from International Classification of Diseases codes and thus may not be accurate. Additionally, the controls in this GWAS were not age-matched with AD cases, which may exacerbate survival bias. Therefore, we are doubtful that this GWAS produced genuine AD associations that can be considered the gold standard. Second, it is unclear what metrics should be used to benchmark the performance of GWAX. In this study, we used the genetic correlations with EA and coronary artery disease to quantify the effectiveness of bias-reduction approaches, but fully addressing issues in GWAX would require replication and functional validation of findings. Third, the nonpresentiveness of UKB participants is well documented[39–41], but it has been suggested that some sampling issues in UKB are not observed in other cohorts[40]. Additionally, we focused only on individuals of European descent. It is an important future direction to investigate how these issues generalize to other ancestries and cohorts.

Taken together, our findings have important implications for the field, as they uncover an urgent, ubiquitous, yet understudied problem hidden in plain sight. Given the popularity of GWAX and the potential of creating misleading results, it is of great importance to reassess the statistical foundation of GWAX. We urge the research community to critically reconsider the applications of family history-based proxy phenotypes and adopt a more cautious and rigorous approach when drawing conclusions from GWAX findings. An immediate remedy for all future studies is to release GWAS and GWAX summary statistics separately for research use, although fully addressing these issues will most likely require tremendous efforts in results validation and development of novel statistical methodologies.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01963-9.

References

1. Abdellaoui, A., Yengo, L., Verweij, K. J. & Visscher, P. M. 15 years of GWAS discovery: realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).
2. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
3. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case–control association mapping by proxy using family history of disease. *Nat. Genet.* **49**, 325–331 (2017).
4. McKhann, G. et al. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939 (1984).
5. Marioni, R. E. et al. GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 99 (2018).
6. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
7. Schwartzentruber, J. et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat. Genet.* **53**, 392–402 (2021).
8. Wightman, D. P. et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet.* **53**, 1276–1282 (2021).
9. Bellenguez, C. et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
10. Sherva, R. et al. African ancestry GWAS of dementia in a large military cohort identifies significant risk loci. *Mol. Psychiatry* **28**, 1293–1302 (2023).
11. Escott-Price, V. & Hardy, J. Genome-wide association studies for Alzheimer's disease: bigger is not always better. *Brain Commun.* **4**, fcac125 (2022).
12. Grotzinger, A. D., Fuente, J., Privé, F., Nivard, M. G. & Tucker-Drob, E. M. Pervasive downward bias in estimates of liability-scale heritability in genome-wide association study meta-analysis: a simple solution. *Biol. Psychiatry* **93**, 29–36 (2023).
13. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
14. Demange, P. A. et al. Investigating the genetic architecture of noncognitive skills using GWAS-by-subtraction. *Nat. Genet.* **53**, 35–44 (2021).
15. Rietveld, C. A. et al. Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc. Natl Acad. Sci. USA* **111**, 13790–13794 (2014).
16. Larsson, S. C. et al. Modifiable pathways in Alzheimer's disease: Mendelian randomisation analysis. *BMJ* **359**, j5375 (2017).
17. Andrews, S. J. et al. Causal associations between modifiable risk factors and the Alzheimer's phenome. *Ann. Neurol.* **89**, 54–65 (2021).
18. Lambert, J.-C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
19. Hujoel, M. L. A., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case–control status and family history of disease increases association power. *Nat. Genet.* **52**, 541–547 (2020).
20. de la Fuente, J., Grotzinger, A. D., Marioni, R. E., Nivard, M. G. & Tucker-Drob, E. M. Integrated analysis of direct and proxy genome wide association studies highlights polygenicity of Alzheimer's disease outside of the *APOE* region. *PLoS Genet.* **18**, e1010208 (2022).

21. Liu, H. et al. Mendelian randomization highlights significant difference and genetic heterogeneity in clinically diagnosed Alzheimer's disease GWAS and self-report proxy phenotype GWAX. *Alzheimer's Res. Ther.* **14**, 17 (2022).
22. European Alzheimer's & Dementia Biobank Mendelian Randomization (EADB-MR) Collaboration. Genetic associations between modifiable risk factors and Alzheimer disease. *JAMA Netw. Open* **6**, e2313734 (2023).
23. Thorp, J. G. et al. Genetic evidence that the causal association of educational attainment with reduced risk of Alzheimer's disease is driven by intelligence. *Neurobiol. Aging* **119**, 127–135 (2022).
24. Chen, Y. et al. Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. *Nat. Genet.* **55**, 44–53 (2023).
25. Grotzinger, A. D. et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).
26. Wu, Y. et al. Estimating genetic nurture with summary statistics of multigenerational genome-wide association studies. *Proc. Natl Acad. Sci. USA* **118**, e2023184118 (2021).
27. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
28. van Rheenen, W. et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).
29. Ferrari, R. et al. Frontotemporal dementia and its subtypes: a genome-wide association study. *Lancet Neurol.* **13**, 686–699 (2014).
30. Chia, R. et al. Genome sequencing analysis identifies new loci associated with Lewy body dementia and provides insights into its genetic architecture. *Nat. Genet.* **53**, 294–303 (2021).
31. Mignogna, G. et al. Patterns of item nonresponse behaviour to survey questionnaires are systematic and associated with genetic loci. *Nat. Hum. Behav.* **7**, 1371–1387 (2023).
32. Schoeler, T. et al. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat. Hum. Behav.* **7**, 1216–1227 (2023).
33. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
34. Phelan, J. C. & Link, B. G. Fundamental cause theory. in *Medical Sociology on the Move*. 105–125 (Springer, 2013).
35. Pedersen, E. M. et al. Accounting for age of onset and family history improves power in genome-wide association studies. *Am. J. Hum. Genet.* **109**, 417–432 (2022).
36. Tublin, J. M., Adelstein, J. M., del Monte, F., Combs, C. K. & Wold, L. E. Getting to the heart of Alzheimer disease. *Circ. Res.* **124**, 142–149 (2019).
37. Stakos, D. A. et al. The Alzheimer's disease amyloid-β hypothesis in cardiovascular aging and disease: JACC Focus Seminar. *J. Am. Coll. Cardiol.* **75**, 952–967 (2020).
38. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
39. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
40. Pirastu, N. et al. Genetic analyses identify widespread sex-differential participation bias. *Nat. Genet.* **53**, 663–671 (2021).
41. Tyrrell, J. et al. Genetic predictors of participation in optional components of UK Biobank. *Nat. Commun.* **12**, 886 (2021).

Methods

Ethics

We analyzed the individual-level data from the UKB and AllofUs cohorts. Our research complies with all relevant ethical regulations. Collection of the UKB data was approved by the Research Ethics Committee of the UKB. All participants of UKB and AllofUs provided written informed consent. We excluded the samples for those who withdrew from UKB from our analyses.

GWAS analysis in UKB

UKB is a large-scale population-based cohort with more than 500,000 participants from across the UK[2]. Between 2006 and 2010, participation invitations were sent to more than 9 million individuals aged between 40 and 69 years, living within a 25-mile radius of any of the 22 UKB assessment centers and registered with a general practitioner of the UK National Health Service[39]. Approximately 5% of the invitees participated in the study, went through a wide range of physical measures and reported detailed information on their sociodemographic, lifestyle, mental and physical health history and family history.

We conducted GWAS in UKB for parental AD history and parental illness awareness using Regenie[42] (v3.2.2) while controlling for sex, year of birth and genotyping array (data field 22000 in UKB) as fixed-effect covariates. Population stratification was accounted for in the ridge regression step of Regenie, which is similar to a linear mixed-model approach without having to compute the genetic relatedness matrix. We excluded participants with conflicting genetically inferred (data field 22001) and self-reported (data field 31) sex, those who withdrew from the study and those who were recommended to be excluded by UKB (data field 22010). Individuals of European ancestry were identified from principal component analysis (data field 22006). We kept only the SNPs with a missing call rate ≤0.01, minor allele frequency ≥0.01 and Hardy–Weinberg equilibrium test P value $\geq 1 \times 10^{-6}$ using PLINK1.9.

Parental AD history (that is, the outcome in AD GWAX) was derived from survey responses to questions regarding the 'illnesses of father' (data field 20107) and 'illness of mother' (data field 20110). Response options included 'Do not know', 'Prefer not to answer', 'None of the above' or one of 12 diseases, including 'Alzheimer's diseases/dementia'. Participants were coded as proxy cases if either parent had AD and as controls if both parents were not affected by AD. Other samples were removed from the analysis. Additionally, participants who self-identified as adopted (data field 1767) were excluded from the study. We identified 47,993 proxy cases and 315,096 controls in the parental AD GWAX.

The parental illness awareness phenotype was derived from UKB data fields 20107 and 20110. Cases were those who selected 'Do not know (group 1)' or 'Do not know (group 2)' for either their father's or mother's illnesses. Controls were those who selected 'None of the above' or any disease in both groups for both their father's and mother's illnesses. Others were excluded from the analysis. A total of 59,471 cases and 339,170 controls were identified.

GWAS analysis in AllofUs

AllofUs is a nationally representative longitudinal cohort in the USA with a goal of recruiting 1 million participants. Detailed descriptions of the study design and data characteristics have been published previously[43,44]. Briefly, enrollment started in May 2018, and adults who are 18 years or older, have the capacity to consent and reside in the USA or a US territory are eligible to enroll. The program seeks to recruit populations that have been understudied in biomedical research. Participants can enroll through either the AllofUs website or a smartphone app. Individuals from under-represented groups who are enrolled in the program are prioritized for physical measurements and biospecimen collection. All participants are required to complete the Basics, Overall Health and Lifestyle survey modules while other survey modules (including Personal and Family Health History, Health Care Access and Utilization, COVID-19 Participant Experience, COPE Minute Survey and Social Determinants of Health) are optional.

We conducted GWAS using AllofUs samples for two phenotypes: participation in the family health history survey and family medical history awareness. The family health history survey is an optional module, and only a subset of AllofUs samples participated in this module. We determined survey participation status by checking whether an individual answered the first question in this module, which reads, "How much do you know about illnesses or health problems for your parents, grandparents, brothers, sisters, and/or children?". This question has four possible response options: (1) none at all; (2) some; (3) a lot; and (4) skip. The GWAS on family medical history awareness was based on the answers to this question. We coded the responses as follows: none = 0; some = 1; a lot = 2. Individuals selecting skip were excluded from the analysis.

For both GWAS in AllofUs, we used independent samples of European descent and adjusted for biological sex, standardized age, square of the standardized age and the top 16 genetic principal components. GWAS was performed using Hail (v0.2) on the whole-genome sequencing data (v7 data release). Genetic ancestry inferred from the principal components and the genetic relatedness between participants were provided in AllofUs. Samples flagged as outliers were excluded from the analysis. We kept only the SNPs with a missing call rate ≤0.01, minor allele frequency ≥0.01 and Hardy–Weinberg equilibrium test P value $\geq 1 \times 10^{-6}$. Sample size for the GWAS on family medical history awareness was 77,579. There were 78,027 cases (participants) and 47,519 controls (nonparticipants) in the GWAS on participation in the family medical history survey.

Measurement error and winner's curse correction

We used Deming regression implemented in the R package mcr (v1.3.3) to correct for measurement errors in SNP effect estimates. We used the mcreg() function and specified the ratio of the error variances to be 42,706/41,679, where 42,706 was the effective sample size (sum of all the effective sample sizes from all contributing cohorts) for the AD GWAS by Kunkle et al.[13] and 41,649 was the effective sample size for the AD GWAX we performed in UKB.

We used the R package WinCurse (v0.0.1) to correct for winner's curse in the GWAS from Kunkle et al.[13]. The adjusted SNP effect size followed formulas in Turley et al.[45].

Heritability and genetic correlation estimation

We used GNOVA[46] (v2.0) to estimate genetic correlations. We corrected GWAS sample overlap in GNOVA if bivariate LDSC[47] (v1.0.0) output an intercept that was significantly different from zero at $P < 0.05$. We used LDSC to estimate heritability.

PRS regression analysis

We evaluated the performance of AD PRS in HRS. The HRS is a nationally representative longitudinal biennial panel consisting of around 42,000 Americans from 26,000 households that began in 1992. A global cognition composite score was derived from a 27-point scale that included the following: (1) an immediate and delayed 10-noun free recall test to measure memory (0 to 20 points); (2) a serial sevens subtraction test to measure working memory (0 to 5 points); and (3) a counting backward test to measure the speed of mental processing (0 to 2 points). Ten waves of data are available, once every 2 years from 2000 to 2018.

We obtained imputed genetic data from a subset of around 15,000 participants whose genetic information was collected between 2006 and 2012 (NIAGADS accession number NG00119.v1). PRS were calculated using two different approaches: PRS-CS[48] (v1.1.0) and clumped significant SNPs in the GWAS from Kunkle et al.[13]. Only overlapping SNPs that existed in all GWAS summary statistics as well as HRS genotype data were used. We used the PRS-CS-auto implementation to estimate SNP posterior effect sizes from the genome-wide

summary statistics. The second PRS approach weighted allele counts with effect sizes obtained from GWAS summary statistics and only included independent SNPs reaching genome-wide significance in the GWAS from Kunkle et al.[13]. Clumping was executed using PLINK1.9, with clumping parameters set at an r^2 value of 0.1 and a window size of 1,000 kb. We also generated an additional set of PRS excluding the *APOE* region by removing all SNPs in the region (chr19: 45116911–46318605; GRCh37).

To analyze longitudinal cognition data in HRS, we used random intercepts in a linear mixed model to account for within-sample (repeated measures) and within-family (related samples) correlations. The regression analyses were performed using the lme4 (v1.1.35.3) package in R, where we regressed the cognitive scores against PRS while controlling for age, age squared, education years, year the respondent entered the study, sex and the top five genetic principal components. Individual and family IDs were coded as random effects. Only HRS participants of European descent were included in the analysis with a total sample size of 12,018.

Causal effect estimation

Mendelian randomization was conducted using the TwoSampleMR (v0.5.6) package in R[49]. To infer the causal effect of EA on AD, we first clumped EA GWAS summary statistics with $r^2 = 0.001$ and window size = 10,000 kb in PLINK1.9 and then selected only those SNPs reaching genome-wide significance ($P < 5 \times 10^{-8}$) as instruments. The mr function was used to estimate causal effects based on the inverse variance weighted approach. Cochran's Q test was used to test for heterogeneity in the fixed-effects inverse variance weighted approach. MR-Egger was used to test for directional pleiotropy. The weighted median approach was used to assess the robustness of the effect estimates against invalid instruments.

GSUB: a new implementation for GWAS-by-subtraction

The GWAS-by-subtraction model, shown in Fig. 3, aimed to subtract genuine AD associations from GWAX associations based on AD family history (that is, decomposing GWAX into AD and non-AD components). Five parameters were estimated (that is, $\lambda_{11,12,22}$ and $\gamma_{1,2}$) and the main parameter of interest (that is, the SNP effect γ_2 on the non-AD component) is highlighted in the figure in red. First, we can write the expressions for AD and AD family history phenotypes in the liability scale:

$$\text{AD} = \lambda_{11} F_{\text{AD}} + u_{\gamma_1}$$

$$= \lambda_{11}\left(\gamma_1 G + u_{F_1}\right) + u_{\gamma_1}$$

$$= \beta_{\text{GWAS}} G + e_1$$

$$\text{AD family history} = \lambda_{22} F_{\text{non}} + \lambda_{12} F_{\text{AD}} + u_{\gamma_2}$$

$$= \lambda_{22}\left(\gamma_2 G + u_{F_2}\right) + \lambda_{12}\left(\gamma_1 G + u_{F_1}\right) + u_{\gamma_2}$$

$$= \beta_{\text{GWAX}} G + e_2$$

The variances and covariances of the genetic components of the two phenotypes are:

$$\text{Var}\left(\lambda_{11} F_{\text{AD}}\right) = \lambda_{11}^2 \text{Var}\left(F_{\text{AD}}\right) \approx \lambda_{11}^2 = h_1^2$$

$$\text{Var}\left(\lambda_{22} F_{\text{non}} + \lambda_{12} F_{\text{AD}}\right) = \lambda_{22}^2 \text{Var}\left(F_{\text{non}}\right) + \lambda_{12}^2 \text{Var}\left(F_{\text{AD}}\right) \approx \lambda_{22}^2 + \lambda_{12}^2 = h_2^2$$

$$\text{Cov}\left(\lambda_{11} F_{\text{AD}}, \lambda_{22} F_{\text{non}} + \lambda_{12} F_{\text{AD}}\right) = \lambda_{11}\lambda_{12} \text{Var}\left(F_{\text{AD}}\right) \approx \lambda_{11}\lambda_{12} = \sigma_{12}$$

Here, G is the SNP allele count and F_{AD} and F_{non} are the two latent factors (with variance of 1) underlying AD and AD family history. β_{GWAS} and β_{GWAX} are the SNP effect sizes in GWAS and GWAX, respectively, and u and e are residuals. From the first two equations, we have:

$$\beta_{\text{GWAS}} = \lambda_{11}\gamma_1$$

$$\beta_{\text{GWAX}} = \lambda_{22}\gamma_2 + \lambda_{12}\gamma_1$$

Based on this, we obtained the expressions for γ_1 and γ_2:

$$\gamma_1 = \frac{\beta_{\text{GWAS}}}{\lambda_{11}}$$

$$\gamma_2 = \frac{\beta_{\text{GWAX}} - \lambda_{12}\gamma_1}{\lambda_{22}} = \frac{\beta_{\text{GWAX}} - \lambda_{12}\beta_{\text{GWAS}}/\lambda_{11}}{\lambda_{22}}$$

From the third to fifth equations, we can solve for the three loading factors:

$$\lambda_{11} = \sqrt{h_1^2}$$

$$\lambda_{12} = \frac{\sigma_{12}}{\lambda_{11}}$$

$$\lambda_{22} = \sqrt{h_2^2 - \lambda_{12}^2} = \sqrt{h_2^2 - \frac{\sigma_{12}^2}{h_1^2}} = \sqrt{h_2^2\left(1 - r_{12}^2\right)}$$

To estimate the five parameters, we plugged in the SNP effect size estimates and their standard errors from the summary statistics, the estimates for LDSC heritability and genetic covariance between the two traits:

$$\hat{\gamma}_1 = \frac{\hat{\beta}_{\text{GWAS}}}{\sqrt{\hat{h}_1^2}}$$

$$\hat{\gamma}_2 = \frac{\hat{\beta}_{\text{GWAX}} - \hat{\beta}_{\text{GWAS}}\hat{\sigma}_{12}/\hat{h}_1^2}{\sqrt{\hat{h}_2^2 - \hat{\sigma}_{12}^2/\hat{h}_1^2}}$$

The standard error for $\hat{\gamma}_1$ was approximated by:

$$\text{s.e.}\left(\hat{\gamma}_1\right) \approx \frac{\text{s.e.}\left(\hat{\beta}_{\text{GWAS}}\right)}{\sqrt{\hat{h}_1^2}}$$

We note that, based on this model setting, GWAS for the AD factor is the input AD case–control GWAS multiplies by a scaling factor.

To obtain the standard errors for $\hat{\gamma}_2$, we need the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$. When there are sample overlaps between GWAS and GWAX, their covariance can be estimated using the intercept from the bivariate LDSC:

$$\text{s.e.}\left(\hat{\gamma}_2\right) \approx \frac{1}{\hat{\lambda}_{22}}\sqrt{\text{s.e.}\left(\hat{\beta}_2\right)^2 + \left(\frac{\hat{\lambda}_{12}}{\hat{\lambda}_{11}}\right)^2 \text{s.e.}\left(\hat{\beta}_1\right)^2 - 2\frac{\hat{\lambda}_{12}}{\hat{\lambda}_{11}}\widehat{\text{gcov}}_{\text{int}}\,\text{s.e.}\left(\hat{\beta}_1\right)\text{s.e.}\left(\hat{\beta}_2\right)}$$

where $\widehat{\text{gcov}}_{\text{int}}$ is the bivariate LDSC intercept.

We note that similar derivations for the point estimate of $\hat{\gamma}_2$ have been previously shown in the supplementary note of Demange et al.[14]. Here, we provide details for the standard error estimation and have implemented the approach as an open-source software.

Simulations

We conducted simulations to compare our analytical approach for GWAS-by-subtraction with GenomicSEM (v0.0.4). We used HapMap 3 SNP genotype data (853,041 SNPs) from independent UKB samples of European descent. We performed simulations for both quantitative traits (n = 200,000 and 100,000) and binary traits (n = 100,000; case proportion = 20% and 10%). Each setting was repeated 100 times.

Following Fig. 3, we first simulated SNP effect sizes on each latent factor from a normal distribution with a mean of 0 and a variance of $1/M$, where M is the number of causal SNPs. The effect sizes were then transformed by dividing $2p(1-p)$, where p is the minor allele frequency of each SNP. The latent factors F_1 and F_2 were computed as $F_1 = \sum_{j=1}^{M} G_j\beta_{1j}$ and $F_2 = \sum_{j=1}^{M} G_j\beta_{2j}$, respectively, where G_j is the allele count (that is, 0, 1 or 2) for the ith SNP. Then, we calculated the observed continuous trait or disease liabilities as follows.

$$Y_1 = \lambda_{11}F_1 + e_1$$

$$Y_2 = \lambda_{12}F_1 + \lambda_{22}F_2 + e_2$$

For binary traits, we set samples at the top 10% or 20% of disease liability as cases and others as controls. In each repeat, we randomly selected 10,000 causal SNPs for each latent factor. We set $\lambda_{11} = \sqrt{0.5}$, $\lambda_{12} = 0.5$ and $\lambda_{22} = \sqrt{0.5}$.

After simulating the phenotype values, we performed GWAS using PLINK2.0 for each phenotype. Then, we applied GWAS-by-subtraction using both GenomicSEM and GSUB to compare the type I error and power. Due to the computational burden of GenomicSEM, we randomly selected 10,000 null SNPs for the type I error calculation in each repeat. Type I error (and power) were calculated as the percentage of null (and causal) SNPs with P values <0.05.

Approaches for bias reduction in GWAX

We explored several strategies to reduce biases in GWAX. To address survival bias, we implemented two approaches. Following Marioni et al.[5], we required both parents to be older than 65 years, which was determined by either current age (data fields 2946 and 1845) or age at death (data fields 1807 and 3526). We also included parental age (either current age or age at death) as a fixed-effect covariate. There were 36,309 cases and 199,969 controls in this GWAX. Following Jansen et al.[6], we created a continuous 'disease load' based on parental AD status, parental age and AD prevalence in the population: each affected parent contributed 1, while each unaffected parent contributed $\min\{(100 - \text{age})/100, 0.32\}$ to the disease load phenotype, where 0.32 is the population prevalence of AD. Those with unknown parental AD status or parental age were excluded from the analysis, resulting in a sample size of 355,501. We performed GWAS on this continuous outcome while controlling for sex, genotyping array, year of birth and assessment center (data field 54).

We used a weighted GWAS approach to account for nonrandom survey participation. Following Schoeler et al.[32], we calculated two sets of sample weights. To obtain the first set of weights, we used 14 variables to train a participation prediction model by comparing participants who did versus did not report parental illnesses. These included five continuous variables: age, body mass index, weight, height and the age at which full-time education was completed, plus nine categorical variables: household size (1–7 or more individuals), sex (male or female), alcohol consumption frequency (never to daily), smoking habits (never, previous or current smoker), employment status (employed, economically inactive, retired or unemployed), income brackets (from <€18,000 to >€100,000), obesity classification (underweight, healthy weight, overweight or obese), general health status (poor, fair or good) and urbanization level (from village/hamlet to urban). We identified 28,179 independent nonreporting individuals (that is, the nonadopted European-ancestry samples that were not included in AD GWAX) for parental AD history. We then randomly sampled the same number of individuals who reported parental illnesses to match with the nonreporting individuals. We used least absolute shrinkage and selection operator (LASSO) regression in the glmnet (v4.1.6) package in R to predict the reporting of parental illnesses. The model included all the main effects and two-way interaction terms, with the shrinkage parameter λ being determined via fivefold cross-validation. We then conducted weighted GWAS on parental AD history with the remaining samples using weighted least squares in R. We used the Huber–White estimator for the variance of the estimates implemented in the sandwich (v3.0.2) package in R. The sampling weights were calculated as $w = (1-p)/p$, where p represents the probability of reporting, predicted through the trained LASSO model. GWAS covariates included sex, year of birth, year of birth squared, genotyping array and the top 20 principal components. Using the code and dataset shared by Schoeler et al.[32], we computed another set of sampling weights by comparing UKB participants with the general UK population and conducted weighted GWAS on parental AD history. In addition, we also explored using GWAS-by-subtraction to remove participation bias, where we regressed out the participation GWAS from the UKB AD GWAX (Supplementary Table 14). The participation GWAS was conducted using the AllofUs samples, which we have described in detail.

We applied two approaches to adjust for reporting bias. The first approach was to include participants who selected 'Do not know' when answering questions about the illnesses of their father or mother in the analysis as controls and then repeat the AD GWAX (n = 47,993 cases and 349,165 controls). In the second approach, we applied GWAS-by-subtraction to regress out the GWAS on parental illness awareness from AD GWAX.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Summary statistics for the AD GWAX are freely available at http://qlu-lab.org/data.html and the GWAS Catalog (parental AD status: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90448001-GCST90449000/GCST90448951/; parental AD status following the approach of Jansen et al.[6]: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90448001-GCST90449000/GCST90448949/; parental AD status following the approach of Marioni et al.[5]: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90448001-GCST90449000/GCST90448950/; parental health awareness in UKB: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90448001-GCST90449000/GCST90448952/; parental health awareness in AllofUs: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90448001-GCST90449000/GCST90448947/; participation in personal and family medical history survey in AllofUs: https://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90448001-GCST90449000/GCST90448948/). The HRS genetic data were accessed through NIAGADS with accession number NG00119.v1. UKB individual-level data used in the present work were obtained under application no. 42148.

Code availability

The code used in this study is available from the following websites: GSUB, GitHub (https://github.com/qlu-lab/GSUB) or Zenodo[50] (https://doi.org/10.5281/zenodo.13845422); PLINK1.9 (https://www.cog-genomics.org/plink2/) and 2.0 (https://www.cog-genomics.org/plink/2.0/); Regenie, https://github.com/rgcgithub/regenie; Hail, https://hail.is/docs/0.2/index.html; GNOVA, https://github.com/qlu-lab/GNOVA-2.0; LDSC, https://github.com/bulik/ldsc; METAL, https://github.com/statgen/METAL; PRS-CS, https://github.com/getian107/PRScs. We also used the following R packages: GenomicSEM

v0.0.4, tidyverse v2.0.0, data.table v1.14.8, mcr v1.3.3, WinCurse v0.0.1, lme4 v1.1.35.3, TwoSampleMR v0.5.6, sandwich v3.0.2 and glmnet v4.1.6.

References

42. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
43. The All of Us Program Investigators. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
44. The All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
45. Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
46. Lu, Q. et al. A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *Am. J. Hum. Genet.* **101**, 939–964 (2017).
47. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
48. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
49. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
50. qlu-lab. qlu-lab/GSUB. *Zenodo* https://doi.org/10.5281/zenodo.13845422 (2024).

Acknowledgements

The authors gratefully acknowledge research support from National Institute on Aging (NIA) grant R21 AG085162 (Q.L.), NIA Center grant P30 AG017266 (Q.L.) and the University of Wisconsin–Madison Office of the Chancellor and the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation (WARF; Q.L.). The All of Us Research Program is supported by the National Institutes of Health (NIH), Office of the Director, Regional Medical Centers: 1 OT2 OD026549, 1 OT2 OD026554, 1 OT2 OD026557, 1 OT2 OD026556, 1 OT2 OD026550, 1 OT2 OD 026552, 1 OT2 OD026553, 1 OT2 OD026548, 1 OT2 OD026551, 1 OT2 OD026555; IAA AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337 and 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants. The Health and Retirement Study (HRS) genetic data were accessed through NIAGADS with accession no. NG00119.v1. These data were collected with financial support from NIH Director's Opportunity for Research awards using American Reinvestment and Recovery Act funds (RC2 AG036495-01, RC4 AG039029-01). With these funds, the HRS has genotyped almost 20,000 respondents who provided DNA samples and signed consent forms in 2006–2012. The HRS data were produced and distributed by the University of Michigan under the directorship of D. R. Weir, with funding from the NIA (NIA U01AG009470). This research was conducted using the UKB Resource under application no. 42148. We acknowledge the participants and investigators of the FinnGen study. We thank the members of the Social Genomics Working Group at the University of Wisconsin–Madison for helpful comments.

Author contributions

Q.L. conceived and designed the study. Y.W. and Z.S. performed the analyses. Y.W. and Q.L. wrote the manuscript. S.D. implemented the software package for GWAS-by-subtraction. Q.Z. performed the GWAS on parental illness awareness in UKB. J.M. assisted with the mathematical derivations for GWAS-by-subtraction. S.M. advised on the genetics of AD and cognition. J.M.F. advised on the HRS cohort and social science issues. All authors revised and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01963-9.

Correspondence and requests for materials should be addressed to Qiongshi Lu.

Peer review information *Nature Genetics* thanks Michelle Lupton and Bjarni Vilhjálmsson for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Qiongshi Lu

Last updated by author(s): Sep 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	PLINK 1.9 & 2.0 REGENIE 3.2.2 Hail 0.2 LDSC 1.0.0 GNOVA 2.0 PRS-CS 1.1.0 GSUB 1.0.0 METAL released on 2018-08-28 GenomicSEM 0.0.4 R package data.table 1.14.8 R package tidyverse 2.0.0 R package mcr 1.3.3 R package WinCurse 0.0.1 R package lme4 1.1.35.3 R package TwoSampleMR 0.5.6 R package sandwich 3.0.2 R package glmnet 4.1.6

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Summary statistics for the AD GWAX are freely available at http://qlu-lab.org/data.html and GWAS Catalog (GCST90448951 for parental AD status; GCST90448949 for parental AD status following Jansen 2019 approach; GCST90448950 for parental AD status following Marioni 2018 approach; GCST90448952 for parental health awareness in UKB; GCST90448947 for parental health awareness in AllofUs; GCST90448948 for participation of personal and family medical history survey in AllofUs). The Health and Retirement Study genetic data were accessed through NIAGADS with accession number NG00119.v1. UKB individual-level data used in the present work were obtained under application no. 42148.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Biological sex was used as regression covariate.
Reporting on race, ethnicity, or other socially relevant groupings	Samples of European ancestry based on genetics were used in analyses since other genetic ancestry groups do not have sufficient number of samples.
Population characteristics	UKB is a large-scale population-based cohort with more than 500,000 participants aged between 37-73 (median age 58) at enrollment from across the UK. The AllofUs research program is a nationally representative longitudinal cohort in the US. Over 1 million participants has been recruited. More then 400,000 participants aged between 18 to over 100 (median age 53) have taken the survey questionnaires and more than 200,000 participants have whole genome sequencing data. Both are volunteer based population cohorts. We used only European ancestry participants in this study.
Recruitment	For UK Biobank: Between 2006 and 2010, participation invitations were sent to more than 9 million individuals aged between 40 and 69 years, lived within a 25-mile radius of any of the 22 UKB assessment centers, and registered with a general practitioner of the UK National Health Service. About 5% of the invitees participated in the study, went through a wide range of physical measures, and reported detailed information on their sociodemographic, lifestyle, mental and physical health history, and family history. For All of Us: Adults who are 18 years and older, have the capacity to consent, and reside in the US or a US territory are eligible to enroll. It seeks to recruit populations that have been understudied in biomedical research. Participants can enroll either through the AllofUs website or a smart phone app. Individuals from underrepresented groups who are enrolled in the program will be prioritized for physical measurements and biospecimen collections.
Ethics oversight	We analyzed the individual-level data from the UKB and AllofUs cohorts. Our research complies with all relevant ethical regulations. Collection of the UK Biobank data was approved by the Research Ethics Committee of the UK Biobank. All participants of UK Biobank and All of Us provided written informed consent. We excluded the samples who withdraw from the UK Biobank in our analyses.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Data used in this study was collected by external sources and used for secondary analysis. Sample size was not pre-determined. Detailed QC process was described in Methods section. A maximum number of 398,641 and 125,546 samples were used from the UK Biobank and AllofUs, respectively. No sample size calculation was performed and we used maximal possible number of samples in our analyses.
Data exclusions	For genetic quality control, samples were excluded for poor DNA quality and genetic variants were excluded using standard metrics such as minor allele frequency, missingness, Hardy-Weinburg equilibrium and imputation quality. For GWAS, samples were also excluded if we could not determine their phenotypes or have missing regression covariates or withdraw from the study.

Replication	All the data for the plots are either included in the Supplementary Tables or shared via GWAS Catalog.
Randomization	Not applicable; this is an observational study
Blinding	Not applicable; this is an observational study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA