European Archives of Psychiatry and Clinical Neuroscience (2023) 273:1359–1368
https://doi.org/10.1007/s00406-022-01527-0

ORIGINAL PAPER

Phenotypic and genetic factors associated with donation of DNA and consent to record linkage for prescription history in the Australian Genetics of Depression Study

Lina Gomez[1] · Santiago Díaz-Torres[2,3] · Lucía Colodro-Conde[4] · Luis M. Garcia-Marin[1,2] · Chloe X. Yap[5] · Enda M. Byrne[5,6] · Loic Yengo[5] · Penelope A. Lind[2,4,8] · Naomi R. Wray[5,7] · Sarah E. Medland[4] · Ian B. Hickie[9] · Michelle K. Lupton[1] · Miguel E. Rentería[1,2] · Nicholas G. Martin[1] · Adrian I. Campos[1,5]

Received: 10 December 2021 / Accepted: 15 November 2022 / Published online: 24 November 2022

Abstract

Samples can be prone to ascertainment and attrition biases. The Australian Genetics of Depression Study is a large publicly recruited cohort ($n = 20{,}689$) established to increase the understanding of depression and antidepressant treatment response. This study investigates differences between participants who donated a saliva sample or agreed to linkage of their records compared to those who did not. We observed that older, male participants with higher education were more likely to donate a saliva sample. Self-reported bipolar disorder, ADHD, panic disorder, PTSD, substance use disorder, and social anxiety disorder were associated with lower odds of donating a saliva sample, whereas anorexia was associated with higher odds of donation. Male and younger participants showed higher odds of agreeing to record linkage. Participants with higher neuroticism scores and those with a history of bipolar disorder were also more likely to agree to record linkage whereas participants with a diagnosis of anorexia were less likely to agree. Increased likelihood of consent was associated with increased genetic susceptibility to anorexia and reduced genetic risk for depression, and schizophrenia. Overall, our results show moderate differences among these subsamples. Most current epidemiological studies do not search for attrition biases at the genetic level. The possibility to do so is a strength of samples such as the AGDS. Our results suggest that analyses can be made more robust by identifying attrition biases both on the phenotypic and genetic level, and either contextualising them as a potential limitation or performing sensitivity analyses adjusting for them.

Keywords Attrition · Bias · Genetics · Polygenic · Epidemiology · Recruitment · Cohort study · Selection bias · PRS · Depression

✉ Adrian I. Campos
adrianisaac.camposgonzalez@uq.edu.au

[1] Genetic Epidemiology Lab, Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia

[2] School of Biomedical Sciences, Faculty of Medicine, The University of Queensland, Brisbane, QLD, Australia

[3] Statistical Genetics Lab, Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia

[4] Psychiatric Genetics Lab, Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia

[5] Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD, Australia

[6] Child Health Research Centre, The University of Queensland, Brisbane, QLD, Australia

[7] Queensland Brain Institute, The University of Queensland, Brisbane, QLD, Australia

[8] School of Biomedical Sciences, Queensland Institute of Technology, Brisbane, QLD, Australia

[9] Brain and Mind Centre, University of Sydney, Camperdown, NSW, Australia

Introduction

The Australian Genetics of Depression Study (AGDS) is a large cohort study including more than 20,000 participants. Recruitment targeted individuals who had been diagnosed or treated for depression [1]. The majority (75%) of the participants in the study are women. The mean age (at the time of recruitment) was 43 years with a standard deviation of 15 years. A high percentage of the participants (95%) have been diagnosed with depression, and similar to what is seen in population studies, 68% of the cohort reported a history of at least one other comorbid mental health diagnosis. The AGDS collected a vast amount of phenotypic data through online questionnaires, as well as a biological specimen for genotyping. A key component of the study was the optional consent to linkage of participants' prescription history with the study data, as well as the option. These datasets enable a variety of novel analyses such as corroboration of medication self-reports as well as health economics and comorbidity analyses. The other key component was obtaining genotype data through a spit sample; this was also optional and independent of whether consent to linkage was provided.

Public recruitment practices can be prone to biases as they indirectly target a portion of the population that consists of individuals that are willing to participate in the study. Voluntary participation could lead to both over or underrepresentation (relative to the population) of exposures [2], and outcomes [3]. When this happens, collider bias, an artificial association between two variables as the product of either adjusting for a covariate that is actually an outcome of the two variables studied [4], or sampling strategies that modify the likelihood of recruiting participants with specific values for the common outcome [5], may occur. This is possible regardless of whether selection occurs for entry into the study, or for the completion of optional modules in subsequent waves of data collection. We refer to this phenomenon as attrition. Previous studies have shown that attrition can influence measures of association [6] by creating a bias in the prevalence and incidence of the variables studied.

Large cohorts and genetic studies are not free of sampling and attrition biases. Evidence of such include the observation of unexpected autosomal heritability of sex in cohorts with active recruitment, which could possibly reflect differences in genetic factors driving participation in males and females [7], and genetic factors associated with participating in optional subsections of surveys of the UK-Biobank [8, 9], all of which may lead to incorrect inferences in downstream analyses. Identifying potential sources of selection bias is necessary to perform statistical adjustments such as propensity score weighting, standardization or matching to a relevant population [10].

A sample with genotype data can, in theory, be compared with a reference sample to identify unmeasured sources of selection. That is, assuming that selection is based on a heritable trait, as well as the existence of a sufficiently powered PRS for said trait. For example, a sample ascertained for individuals with higher education should have a higher mean educational attainment (EA) PRS when compared to a random sample from the same population. Assuming EA was not measured in this sample, EA PGS could be compared to a population sample to identify ascertainment and standardize or weight to adjust for potential biases. We believe this strategy may be possible when coupled with standardization, because we expect a well powered PRS to serve as a proxy for the unobserved trait and expect any ancestral/population selection in PRSs to be relatively small. Thus, genetics holds promise to enable assessment of selection factors provided a sufficiently good PRS can be estimated, and a reference dataset is available. It remains an open question whether this theoretically sound approach will be practically useful.

The objective of this study was to evaluate selection biases in two aspects of participation in the ADGS: donating a saliva sample for genotyping and agreeing to linkage to medical prescription records (PBS linkage) in the AGDS. We did this by investigating sociodemographic and psychiatric differences between participants who donated a saliva sample compared to those who did not. We further tested whether there is evidence of a heritable component underlying agreement to PBS linkage, as it would evidence whether any genetic correlate are expected. We also investigated sociodemographic, clinical, and polygenic differences between participants who agreed to PBS linkage and those who did not. As part of this study, we aim to investigate if genetic data could be used to identify sources of selection had these been not observed and to exemplify potential biases of downstream analyses using simulations based on the AGDS data. Finally, we compare our findings across the two participation measures and with studies of participation in other cohorts such as the UK-Biobank to assess whether the same factors underlie participation.

Methods

Sample recruitment

The Australian Genetics of Depression Study recruited 22,424 Australian participants through two avenues: a mail-out to patients who had at least 4 prescriptions for antidepressants in the previous 5 years (14%) and a media campaign to recruit patients who had received a diagnosis of depression from a doctor, psychiatrist or psychologist (86%). Potential participants were directed to the AGDS website (https://www.geneticsofdepression.org.au), and

informed consent was gathered prior to data collection through online questionnaires. The AGDS inclusion criteria included (1) reporting that they had been treated for depression by a health professional (2) agreeing to donate a saliva sample for genotyping (although only 72.5% actually did so). Full details for the AGDS can be found elsewhere [1].

Among the phenotypes collected, participants confirmed whether they had taken any of the ten most commonly prescribed antidepressants in Australia. For each antidepressant taken, we gathered data on antidepressant efficacy and experienced side effects. Data on demographics, clinical history of psychiatric disorders as well as personality traits (neuroticism and extraversion scores) were also collected. Furthermore, willing participants provided optional consent to record linkage of their Pharmaceutical Benefits Scheme (PBS) prescription and Medical Benefits Scheme (MBS) records, which could be used to validate medication self-reported data, assess concurrent medications, and infer comorbidities.

In this study, we focus on (1) whether participants agreed to record linkage of their PBS records for research purposes as the outcome of interest, and (2) whether they donate a saliva sample for DNA genotyping. The full list and details of instruments used for AGDS phenotyping are available at https://bit.ly/3y72lyg. All participants provided informed consent prior to participating in the study. The QIMR Berghofer Medical Research Institute Human Research Ethics Committee approved all questionnaires and research procedures for the AGDS under project number P2118.

Genotyping imputation and quality control

Upon completion of the core questionnaire, participants were mailed a GeneFix GFX-02 2 mL saliva DNA extraction kit (Isohelix plc) to use at home and then returned by mail for subsequent genotyping. The AGDS sample was genotyped using the Illumina Global Screening Array (GSA V.2.0). Genotype data were cleaned by removing unknown or ambiguous map position, strand alignment, high missingness (> 5%), deviation from Hardy–Weinberg equilibrium, low minor allele frequency (< 1%), and GenTrain score < 0.6 variants. Imputation was performed through the Michigan imputation server web service using the HRCr1.1 reference panel, as the majority of the cohort were of European ancestry. Genotyped individuals were excluded from polygenic risk score (PRS) analyses based on high genotype missingness, inconsistent and unresolvable sex, or if deemed ancestry outliers from the European population, based on principal components derived from the 1000Genomes reference panel (defined as > 6SD from the PC1/PC2 centroid).

SNP-based heritability

We employed genome-based restricted maximum likelihood (GREML) as implemented in GCTAv.1.91.7 [11] to estimate the proportion of variance in consent to record linkage (on the observed scale) explained by measured genetic differences (*SNP-based heritability*). This approach leverages a genetic-relatedness matrix (GRM) and restricted maximum likelihood to partition the variance of a phenotype into a genetic and environmental component [12]. The underlying logic is to assess whether genetic covariation explains a significant proportion of the covariation of a trait. If so, then we expect the ascertained trait can be identified using genetic data. In a large sample, a GWAS of participation followed by genetic correlation analyses could be performed. Nonetheless, for the sample size of the AGDS this strategy is not feasible, but PRS could potentially help in identifying these traits. We note that this strategy will only work if the association between selection and the PGS is driven through the PGS (focal) phenotype; associations for other reasons could fail to identify ascertainment or could even induce its own confounding if adjusted for. For this study, a GRM based on a subset of unrelated (genomic relatedness cut-off < 0.05) individuals of European ancestry was employed to identify evidence for a genetic component to consenting to record linkage of PBS records.

Polygenic risk scores (PRS)

We computed PRS in order to test to what extent the genetic risk for traits that showed a phenotypic association or have previously been linked to attrition in other studies (see "Discussion" section), predicted consent to PBS record linkage. For the PRS predictions, we excluded European ancestry outliers and used only one member from groups of related individuals (genomic relatedness cut-off < 0.05) to avoid confounding from cryptic relatedness, as this violates the independence assumption of classical logistic regression. We estimated PRS for educational attainment [13], neuroticism [14], major depressive disorder [15], bipolar disorder [16], schizophrenia (SCZ) [17], and anorexia nervosa [18] using GWAS summary statistics without sample overlap with the AGDS cohort. SBayesR was used to estimate the joint GWAS effect sizes adjusting for the correlation between SNPs [19]. Prior to estimating PRS, we excluded low imputation quality ($r^2 < 0.6$), MAF < 0.01, non-autosomal, and strand-ambiguous variants. Imputed genotype dosage data were used to calculate PRS by multiplying the variant effect size by the dosage of the effect allele. Finally, the total sum was calculated across all variants. This procedure was performed using Plink 1.9 [20]. A relevant follow-up question is how much of the *SNP-based heritability* is explained jointly by these genetic factors. To estimate the

proportion of *SNP-based heritability* explained by these PRS and whether there is any residual genetic variance for selection, a secondary GREML analysis including all nominally significant PRS (SZC, MDD, and anorexia) as a fixed effect was performed, and the SNP-based heritability of both models (with and without PRS) were compared using the formula

$$1 - \frac{VG_{adj}}{VG_0},$$

where VG_0 is the genetic variance component estimated in the null model (i.e., without PRS) and VG_{adj} is the same estimate but of the model including PRS for SCZ, MDD and anorexia as fixed effects. This approach is likely to yield an underestimate as the GRM used to estimate SNP-based heritability is constructed using SNPs also included in the PRS.

Illustrating-biased associations in the AGDS

To illustrate a possible scenario whereby selection induces bias, we simulated a collider variable for anorexia. That is, a variable that has no real association with anorexia, but would show a spurious association upon conditioning on agreeing to PBS linkage. We can envision such a scenario if we are performing complete data analysis of a variable obtained from PBS records. These analyses would inherently be stratified (i.e. focus only on participants that agreed to linkage). We simulated a variable (named collider here) influencing agreeing to PBS linkage independently from anorexia. We varied the effect size of the collider on agreeing to PBS linkage, covering the range of selection effect sizes observed by our study (OR range from 1.002 to > 1.8), conditional on anorexia to ensure independence, and assess whether we identify a spurious association between anorexia and the collider. The actual effect of the collider on agreeing to PBS linkage was determined using a logistic regression. This simulation was repeated 100 times for each effect size and each type of analysis (see below). Finally, we used post-stratification matching [10] based on rates of anorexia and average of the collider to show how the bias can be alleviated. Simulations were performed based on the AGDS data and using the *R* and the *MatchIt* library. The types of analyses performed were *standard*—negative control using all data and not conditioning on agreeing to linkage; *conditional*—as positive control testing for association between anorexia and collider while conditioning on agreeing to linkage; *stratified*—performing the association test only within participants that agreed to linkage; *stratified matched* analysis—where the sample that agreed to linkage is post-stratified to match the whole sample in terms of prevalence of anorexia and mean collider; *stratified subsampled*—same as stratified but ensuring the same sample size as that of the *stratified matched* subsamples (to ensure results are not due to power differences).

Statistical analyses

Logistic regression was used to examine the association between our outcomes (i.e., consent to record linkage and providing a saliva sample) and variables of interest including age, sex, educational attainment, psychopathology, and PRS. The non-genetic regressions were adjusted for sex and age at study enrolment. Sensitivity regressions using genetic variables including age, sex, and the first 20 genetic principal components were performed to test whether associations with participation may be explained by the relationship of selection on age, sex and population stratification. Results assessing the proportion of *SNP-based heritability* used linear mixed-effects models fit via restricted maximum likelihood [12]. Nominally significant results are defined as those with $p < 0.05$ and statistical significance was defined after Bonferroni correction for multiple testing.

Results

Demographic factors and samples description

Table 1 shows demographic information across AGDS participants who did or did not consent to record linkage. We also contrast demographic factors between participants who donated a saliva sample for genotyping and those who did not. Participants agreeing to record linkage consent were, on average, older (OR = 1.016; 95% CI [1.01–1.02] per year of age). Female participants were less likely to provide consent for record linkage (OR = 0.69; 95% CI [0.63–0.74]). Overall, the cohort's educational attainment is high (e.g., ~ 25% of participants reported having a postgraduate degree). We did not observe a significant association between educational attainment and providing consent for record linkage (OR = 1.02; 95% CI 0.99–1.05). Being male (OR = 1.13; 95% CI [1.05–1.23]), older (OR = 1.012; 95% CI [1.010–1.015] per year of age), and having higher educational attainment (OR = 1.18; 95% CI [1.14–1.21] per educational category) were all associated with donating a saliva sample for genotyping.

Associations with psychiatric and personality traits

The association between self-reported lifetime psychiatric diagnosis and differential consent for record linkage or biological sample are shown in Table 2. Participants with self-reported diagnosis of bipolar disorder (OR = 1.31 95% CI [1.17–1.46]) were more likely to consent to record linkage. In contrast, participants who reported a diagnosis

Table 1 Relationship between demographics and consent for record linkage and provision of a genetic sample

	Whole sample	Record linkage	No record linkage	Saliva sample	No Saliva sample
Age in years					
Mean (SD)	43.17 (15.42)	44.3 (15.26)	39.95 (15.42)	43.7 (15.3)	41.9 (15.7)
Sex					
Female	16,790 (75%)	12,072 (72%)	4718 (28%)	12,166 (72.5%)	4624 (27.5%)
Male	5474 (24.5%)	4383 (80%)	1091 (20%)	3937 (71.9%)	1537 (28.1%)
Education					
No formal education	12 (0.06%)	7 (0.05%)	5 (0.01%)	7 (0.04%)	5 (0.11%)
Primary school	53 (0.3%)	44 (0.3%)	9 (0.2%)	38 (0.24%)	15 (0.36%)
Junior secondary school	1162 (5.75%)	893 (5.9%)	269 (5.3%)	862 (5.4%)	300 (7.2%)
Senior secondary school	1714 (8.5%)	1232 (8.1%)	482 (9.6%)	1252 (7.8%)	462 (11.1%)
Certificate or diploma	4793 (23.7%)	3692 (24.3%)	1101 (21.9%)	3755 (23.4%)	1038 (24.9%)
Degree	7055 (34.9%)	5206 (34.3%)	1849 (36.8%)	5617 (35.1%)	1438 (34.5%)
Postgraduate degree	5399 (26.7%)	4088 (27.0%)	1311 (26.1%)	4485 (28.0%)	914 (21.9%)

Table 2 Association between psychiatric traits and participation

Disorders	Consent to record linkage OR (95% CI)	Consent to record linkage *p*-value	Saliva sample OR (95% CI)	Saliva sample *p*-value	*N* cases
Bipolar disorder	1.31 (1.17–1.46)	1.51e−06[a]	0.85 (0.77–0.94)	0.001[a]	2057
Anxiety disorder	0.92 (0.87–0.98)	0.007	0.92 (0.87–0.98)	0.007	12,131
Schizophrenia	1.22 (0.87–1.7)	0.255	0.78 (0.58–1.05)	0.100	195
ADHD	0.98 (0.85–1.14)	0.828	0.71 (0.62–0.82)	2.37e−06[a]	900
Agoraphobia	0.92 (0.76–1.13)	0.437	0.81 (0.67–0.98)	0.03	488
Premenstrual dysphoric disorder	0.9 (0.74–1.09)	0.292	0.87 (0.72–1.06)	0.17	504
Anorexia nervosa	0.57 (0.5–0.64)	3.97e−19[a]	1.76 (1.51–2.05)	3.92e−13[a]	1135
Bulimia nervosa	0.7 (0.6–0.81)	4.06e−06[a]	1.16 (0.98–1.36)	0.083	771
Autism spectrum disorder	0.92 (0.73–1.16)	0.468	0.85 (0.68–1.06)	0.149	356
Panic disorder	0.89 (0.81–0.99)	0.027	0.77 (0.7–0.85)	1.39e−07[a]	2092
Obsessive compulsive disorder	0.81 (0.72–0.92)	0.001[a]	0.93 (0.82–1.05)	0.253	1277
PTSD	0.93 (0.85–1.01)	0.087	0.83 (0.76–0.9)	6.63e−06[a]	3130
Personality disorder	1.02 (0.89–1.15)	0.816	0.83 (0.74–0.94)	0.003[a]	1284
Substance use disorder	1.32 (1.11–1.56)	0.001[a]	0.79 (0.68–0.91)	0.001[a]	829
Social anxiety disorder	0.86 (0.78–0.94)	0.001[a]	0.82 (0.75–0.89)	1.18e−05[a]	2483
Seasonal affective disorder	0.99 (0.82–1.18)	0.881	0.88 (0.74–1.05)	0.161	629
Neuroticism[b]	0.99 (0.98–1.01)	0.310	0.98 (0.97–0.99)	0.002	NA
Extraversion[b]	0.99 (0.98–1)	0.051	1 (0.99–1.01)	0.575	NA

[a] $p < 0.0028$ significant after multiple testing correction

[b] Effect per unit of raw score. Results adjusting for age and sex are available in Supplementary Table S1

of anorexia (OR = 0.57 95% CI [0.50–0.64]) were less likely to consent to record linkage. There was a nominally significant association between consent to record linkage and participants reporting a lifetime diagnosis of substance use disorder (OR = 1.32 95% CI [1.11–1.56] or anxiety disorder (OR = 0.92 95% CI [0.87–0.98]). For the outcome of donating a saliva sample, the following were associated with lower odds of donation, bipolar disorder (OR = 0.85 95% CI [0.77–0.94]), ADHD (OR = 0.71 95% CI [0.62–0.82]), panic disorder (OR = 0.77 95% CI [0.7–0.85]), PTSD (OR = 0.83 95% CI [0.76–0.9]), substance use disorder (OR = 0.79 95% CI [0.68–0.91]) and social anxiety disorder (OR = 0.82 95% CI [0.75–0.89]). Conversely, participants reporting a lifetime diagnosis of anorexia showed higher odds (OR = 1.84 95% CI [01.51–2.05]) of providing a saliva sample (Fig. 1). Most of these associations remained similar regardless of adjustment for age and sex. However, neuroticism became

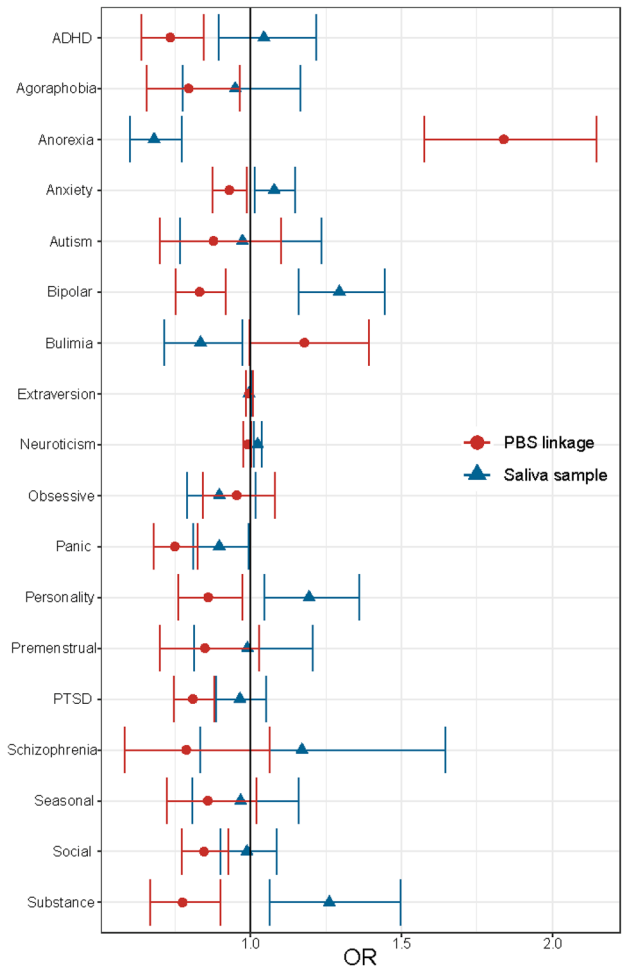

Fig. 1 Associations with neuropsychiatric traits. Forest plots depict odds ratios (OR) and 95% confidence intervals for the association between neuropsychiatric traits and, donating a saliva sample (blue triangular markers) or agreeing to pharmaceutical benefits scheme (PBS) linkage (red circular markers) adjusting for age and sex

positively associated with consent to record linkage after adjustment (Supplementary Table S1).

Genetic factors

This section focuses on consenting to record linkage amongst those who provided a DNA sample. Necessarily, genetic analyses comparing participants who did or did not provide a biological specimen are not possible. Amongst those who provided a biological sample, 79% consented to PBS, implying a binomial variance of 0.79 (1–0.79) = 0.17 on the observed scale. The GREML analysis suggested the presence of a genetic contribution to the likelihood of consenting to record linkage. The *SNP-based* heritability on the observed scale was 0.12 (SE = 0.03, $p = 1.1e{-}7$, phenotypic variance on the observed scale ~ 0.16). We hypothesised that the genetic risk (operationalised as PRS) for the psychiatric traits identified above (e.g., bipolar disorder, anorexia, or neuroticism) would be associated with differential consent for record linkage.

We used logistic regressions to test for association between agreeing to PBS linkage and PRS for neuroticism, MDD, SCZ, or EA. PRS were validated by first predicting the specific trait of interest. All PRS were predictive of their respective traits. For example, neuroticism PRS was strongly associated with neuroticism score. Although statistically significant, SCZ PRS had the lowest evidence for association with its cognate trait ($p = 2.0e{-}4$; Table 3). Anorexia PRS (OR = 0.93 95% CI [0.89–0.97]) and SCZ PRS (OR = 0.91 95% CI [0.86–0.96]) were associated with lower odds of consenting to linkage. However, the latter was not significant after adjusting for genetic principal components (Supplementary Table S2). Conversely, MDD PRS was associated with higher odds of consent to record linkage (OR = 1.07 95% CI [1.02–1.11]; Table 3). Using a GREML analysis, we estimated that MDD, SCZ, and anorexia PRS accounted

Table 3 PRS validation and results of association with consent to record linkage

PRS	Validation association[c]				Association with consent to record linkage		
	Variance explained (R^2)	Effect size[a]	Standard error	*p* value	Variance explained (R^2)	OR (95% CI)	*p* value
Major depression	–	–	–	–	0.001	1.07 (1.02–1.11)	1e−3[b]
Bipolar disorder	0.005	0.18	0.03	2.0e−09[b]	1.06e−5	1.00 (0.97–1.05)	0.75
Educational attainment	0.058	0.28	0.01	1.7e−195[b]	1.23e−5	1.00 (0.97–1.05)	0.74
Schizophrenia	0.010	0.47	0.13	2.0e−4[b]	0.001	0.91 (0.86–0.96)	8e−4[b]
Neuroticism	0.018	0.44	0.03	6.2e−51[b]	3.5e−4	1.03 (0.99–1.08)	0.06
Anorexia	0.010	0.27	0.04	8.2e−13[b]	0.001	0.93 (0.89–0.97)	8e−4[b]

[a]Effect sizes (beta or log of the odds ratio) per standard deviation of PRS

[b]$p < 0.0084$ (bonferroni corrected threshold)

[c]Validation association corresponds to logistic or linear regressions predicting a trait using the PRS of that specific trait (e.g. bipolar disorder PRS predicting bipolar disorder lifetime diagnosis). R^2 for binary traits is calculated as the Nagelkerke pseudo R^2

for around 12% of the SNP-based heritability of consent to record linkage (see "Methods" section) which would account for around 1.4% of the total phenotypic variance on the observed scale. Notably, all PRS were still significantly associated with agreeing to record linkage when jointly estimating their effects using GREML (MDD PRS beta = 0.012 SE = 3.7e−3; SCZ PRS beta = −0.013 SE = 4.9e−3; anorexia PRS = −9.2e−3 SE = 3.8e−3).

Simulation of a collider for anorexia

We simulated a *collider* variable that increases the odds of a participant agreeing to PBS consent. This would induce a spurious association with anorexia (see "Methods" section and Fig. 2a). As expected, conditional analyses are more likely to detect the spurious association compared to a stratified analysis. Nonetheless, stratified analyses should be considered more realistic given that they represent studies using the PBS data available. Our results suggest that small effects on agreeing to PBS linkage (i.e. standardized odds ratio of less than 1.03) would result in undetectable collider bias with anorexia in conditional and stratified analyses. For effect sizes larger than that, a spurious association was identified more often than expected by chance (Fig. 2b). For reference, factors such as bipolar and obsessive–compulsive disorder showed stronger selection effects (ORs = 0.81 and 1.31, respectively) than those required to induce collider with anorexia. We performed follow-up analyses searching for evidence of collider bias with consent. Briefly, we performed association analyses between pairs of traits in either subsample (PBS or saliva donation) and in the whole sample. If an association was identified in the subsample but not in the whole sample this could be indicative of collider. We found no evidence of such pattern across all pairs of psychiatric traits. Finally, post-stratifying the subsample that agreed to PBS linkage to match the whole AGDS sample (in terms of prevalence of anorexia and mean collider) alleviated the bias within the simulations. This alleviation was not due to a reduction in sample size as an analysis subsampled to the same size as the matched analysis still showed inflation (Fig. 2b).

Discussion

This study investigated systematic differences between participants who consented to record linkage of prescription history and those who did not. We also investigated sociodemographic differences associated with the decision to donate a saliva sample for DNA analysis. Our study is motivated by the need to acknowledge and identify differences that could limit the generalisability of findings from prescription history data, and potentially lead to biased or spurious associations. Nonetheless, it is important to reiterate that

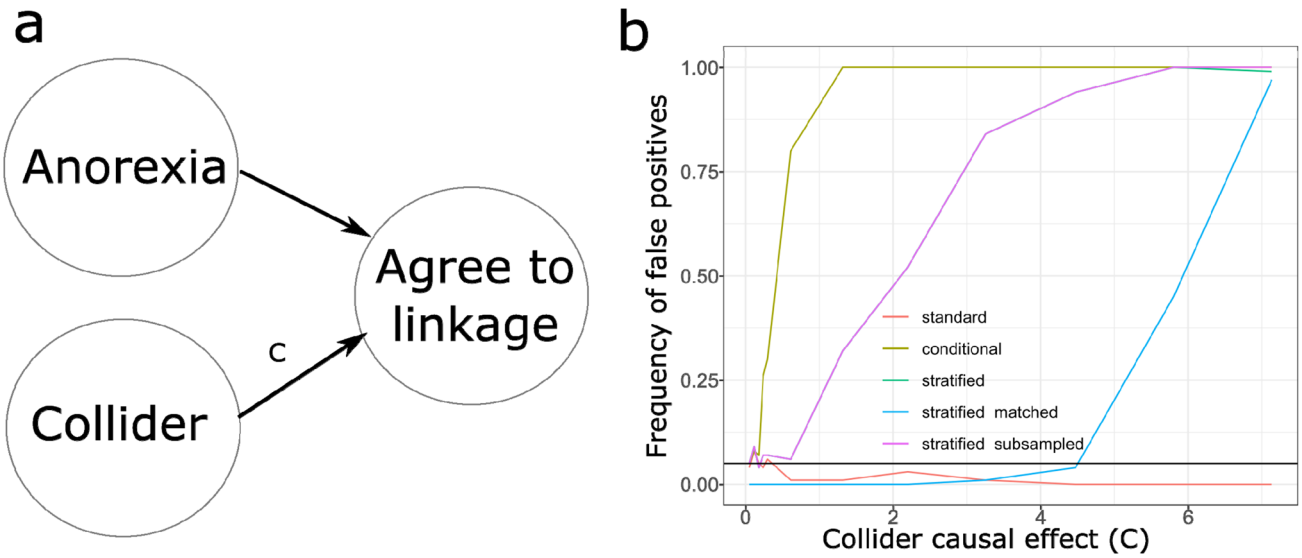

Fig. 2 Simulation illustrates potential collider bias with anorexia. **a** Diagram showing the simulated collider variable exerting a *causal* effect (C) on agreeing to PBS linkage. Note that the effect of the collider is independent of anorexia and we do not expect a significant association between anorexia and collider. **b** Results of 100 simulations for varying levels of collider causal effects. Standard-negative control using all data. Conditional analysis—as positive control testing for association between anorexia and collider while conditioning on agreeing to linkage. Stratified analysis—performing the association test only within participants that agreed to linkage. Stratified matched analysis—where the sample that agreed to linkage is post-stratified to match, in terms of prevalence of anorexia and mean collider as the whole sample. Stratified subsampled—same as stratified but ensuring the same sample size as that of the stratified match subsamples. The black horizontal line depicts a 0.05 false positive rate

representativeness is not necessarily the issue that needs addressing. As previously discussed [21], selection will induce bias if the variables being studied are causally associated with selection.

We illustrate a potential bias simulating a collider to anorexia. Our simulations suggested the causal effect of a *collider* (on selection) needs to be moderate to high to be detectable in the subsample with PBS data and that post-stratification matching can largely alleviate the collider bias. The analyses using matched data inherently had a lower sample size, which could explain non-significant associations by mere power reduction. We showed this was not likely the case as simulations of a conditional analysis subsampled to the same size as the matched dataset, still showed significant inflation. Although many of the identified associations were strong enough to induce collider according to our simulation, it is important to mention that we did not identify evidence of biased associations between pairs of variables that were associated with consent to record linkage or providing a saliva sample.

We identified factors correlated with participation; future studies using AGDS data should take additional caution when performing analyses including these traits. For example, if we were to observe a significant correlation between anorexia PRS and MDD PRS (both associated with participation) in the subset of the sample with PBS data (implicitly conditioning con participation), it would be necessary to perform a sensitivity analysis using the whole sample to assess whether such an association is due to collider. Below, we cover some of the factors identified herein and discuss whether these are consistent with other studies.

Educational attainment has been reported to be a relevant factor for voluntary participation in medical research [9]. In fact, the AGDS has been shown to be a highly educated sample compared to the Australian population [1]. Here, we identified a moderate association between educational attainment and donating a biological sample, but the relationship between educational level and consent to record linkage did not reach statistical significance.

A recent study analysed genetic and demographic factors related to whether UK-Biobank participants shared their email for a follow-up questionnaire. Women and individuals at high genetic risk for schizophrenia were less likely to share their email [8]. While we also identified that females were less likely to consent to record linkage, we found no association between consent to record linkage and self-reported schizophrenia diagnosis. Nonetheless, we found a significant inverse association between SCZ PRS and consent to record linkage, which would align with the findings in the UK-Biobank and suggest our lack of phenotypic association to be due to lack of power from the low number of participants reporting schizophrenia or due to potential unreliability of a self-reported measure. Increased genetic risk for schizophrenia is likely related to higher basal suspiciousness, a heritable trait related to psychiatric disorders [22] which likely leads to a lower desire to agree to record linkage. However, the association between SCZ PRS and agreeing to PBS linkage did not reach statistical significance after adjusting for PCs. This could imply that residual stratification underlies this association, but these results are also consistent with SCZ PRS being the least powered PRS in our study.

Similar research has proposed that participants with higher neuroticism scores and depression are less likely to participate in follow-up questionnaires. [9]. Participants with these characteristics may be more likely to experience feelings of anger, anxiety, and irritability [23] when prompted with follow-up research surveys. In our study, individuals with higher neuroticism scores were more likely to consent to record linkage of their prescription history. We also report an association between neuroticism PRS and neuroticism score. However, the neuroticism PRS was not statistically associated with consent for record linkage. A similar result (i.e. phenotypic association, but no association between PRS and consent to record linkage) for Bipolar disorder was observed. This observation could be explained by the fact that these PRS do not capture all of the heritability for their traits and thus suffer from reduced power, but it may also imply that the relationship between these factors and consent to record linkage is mediated solely through non-genetic factors. We believe the latter scenario to be unlikely.

Participants reporting a lifetime diagnosis of bipolar disorder were more likely to consent to record linkage of their prescription history and less likely to provide a saliva sample for genotyping. Furthermore, genetic risk for depression was nominally associated with greater odds of consenting for record linkage, which is the reverse direction to the previous reports [9]. Participants who reported anorexia were less likely to consent to record linkage, but more likely to provide a biological specimen. This finding may be related to personality characteristics linked to anorexia, such as behaviours in relation to losing control [24]; consenting to access to the clinical prescription history could be violating the concept of having control [25]. However, this explanation fails to explain the positive association with donating a biological specimen.

The lack of consistency between demographic and clinical factors associated with consent for record linkage and donating a saliva sample, coupled with the results of other studies of participation and attrition, leads us to hypothesise that there is no single set of factors underlying attrition to different designs and settings. Evidence of differential factors in our study include the fact that previously reported negative predictors of participation, such as neuroticism and genetic risk for depression [9], were positively associated with consent to record linkage; that educational attainment

was associated with donating a saliva sample, but not with consent to record linkage and that a lifetime diagnosis of anorexia and substance use disorder were associated with both record linkage and saliva donation but with opposing effects. Overall, our results suggest participation biases to be specific to the design and nature of the study or follow-up. That is, there is no *one-size fits all* set of factors systematically driving participation across studies, but the possibility of a set of common factors cannot be ruled out.

It is important to highlight some limitations of this study. The diagnoses studied relied on participant self-report, which could lower the reliability of these diagnoses; however, as mental health disorders are highly impactful, a participants' likelihood to accurately report the diagnoses communicated to them is high. Additionally, participants were informed of the period that would be covered by the record linkage and it is possible that participants who had not experienced a depressive episode in this timeframe were less likely to consent for record linkage (the previous 4.5 years from recruitment). It is also important to note that only Australian citizens and permanent residents have access to the PBS system and we did not collect information on legal citizenship or resident status. Military personnel and their families do not use the MBS/PBS system. As such, non-residents and military personnel (and their families) would not be able to consent to PBS record linkage. The lack of evidence for association between a trait and participation does not necessarily imply a lack of association. This is why we mention and report on nominally significant findings as well. Within the genetic analyses, we excluded participants of non-European ancestry to avoid population stratification. For this reason, care should be taken when generalising these findings to populations with other ancestral backgrounds even within this sample.

Our results show that phenotypic differences can be identified through differences in PRS. The effectiveness of this approach will depend on the heritability of the trait upon which selection took place, as well as the power of the PRS. There are multiple methods for deriving weights for PRS, although they are comparable, some methods may work better for specific traits depending on their genetic architecture. It is also important to acknowledge that PRS themselves could be affected by residual population stratification and selection factors [26, 27]. Thus, a difference between PGS in a sample and a population-based reference may be reflective of either (1) a true correlation between the phenotype and participation (which is also detectable on the genetic level) or (2) residual stratification in the PGS, when it also differentiates the sample from the population. We believe the former should make more sense as PRS gain power and new approaches to adjust for residual biases are developed. Using genetics to estimate and adjust for unobserved biases would be subject to important challenges such as the need for a highly powered PGS and the availability of a population reference sample. PRS may become a useful tool against attrition biases, but further developments are needed to achieve this.

Longitudinal studies with genotype data, such as the AGDS, enable us to identify these biases. Future studies leveraging the prescription data should consider these differences, and perform sensitivity analyses assessing whether their findings could be attributed to the traits identified herein. Most current epidemiological studies do not adjust nor search for attrition biases, let alone at the genetic level. The possibility to do so is a strength of samples such as the AGDS. An example of such an approach is a recent study identifying an unexpected positive genetic correlation between depression and cognitive traits. Follow-up analyses suggested these results to be explained by the fact that the AGDS sample is also highly educated whereas the education levels of the controls were more concordant with the Australian population [28]. We argue that future analyses can be made more robust by identifying biases both on the phenotypic and genetic level, and contextualising them as a potential limitation or performing sensitivity analyses adjusting for them.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00406-022-01527-0.

Acknowledgements Data collection for AGDS was possible thanks to funding from the Australian National Health & Medical Research Council (NHMRC) to NGM, NRW, SEM, IHB, EMB, PAL (GNT1086683) and Medical Research Future Fund (APP1200644). We thank our colleagues Richard Parker, Simone Cross, Scott Gordon and Lenore Sullivan for their valuable work coordinating all the administrative and operational aspects of the AGDS project. NRW thanks the support of NHMRC through Grants 1113400 and 1173790. MER thanks the support of NHMRC and the Australian Research Council (ARC) through an NHMRC-ARC Dementia Research Development Fellowship (GNT1102821). SEM is supported in part by NHMRC investigator Grant APP1172917. The views expressed are those of the authors and not necessarily those of the affiliated or funding institutions.

Author contributions AIC and NGM designed the study. LG, SDT, LMGM, and CY performed the analyses with supervision and guidance from AIC. EMB, PAL, NRW, SEM, IBH, and NGM implemented and supervised the data collection for the AGDS. LCC, LY, and MKL critically appraised the manuscript and gave input in the design and implementation of the analyses. LG, AIC, and MER contributed to the first draft of the manuscript and all authors contributed to editing and drafting the manuscript prior to submission.

Data availability Summary statistics used for PGS are publicly available as described in their respective publications (see methods). Summaries of associations are provided with this manuscript in the main text or supplementary materials. Access to the AGDS data is restricted due to the ethical guidelines governing the study, but may be accessible following ethical review and data transfer agreements, please contact Nicholas Martin (nick.martin@qimrberghofer.edu.au) with any queries related to accessing AGDS data.

Declarations

Competing interests Professor Ian Hickie is the Co-Director, Health and Policy at the Brain and Mind Centre (BMC) University of Sydney, Australia. The BMC operates an early-intervention youth services at Camperdown under contract to headspace. Professor Hickie has previously led community-based and pharmaceutical industry-supported (Wyeth, Eli Lily, Servier, Pfizer, AstraZeneca) projects focused on the identification and better management of anxiety and depression. He is the Chief Scientific Advisor to, and a 5% equity shareholder in, InnoWell Pty Ltd. InnoWell was formed by the University of Sydney (45% equity) and PwC (Australia; 45% equity) to deliver the $30 M Australian Government-funded Project Synergy (2017-20) and to lead transformation of mental health services internationally through the use of innovative technologies. LG, SDT, LCC, LMGM, CY, EMB, LY, PAL, NRW, SEM, IBH, MKL, MER, NGM and AIC have nothing to disclose.

References

1. Byrne EM et al (2020) Cohort profile: the Australian genetics of depression study. BMJ Open 10(5):e032580
2. Oswald LM et al (2013) Volunteerism and self-selection bias in human positron emission tomography neuroimaging research. Brain Imaging Behav 7(2):163–176
3. Patten SB (2000) Selection bias in studies of major depression using clinical subjects. J Clin Epidemiol 53(4):351–357
4. Day FR et al (2016) A robust example of collider bias in a genetic association study. Am J Hum Genet 98(2):392–393
5. Griffith GJ et al (2020) Collider bias undermines our understanding of COVID-19 disease risk and severity. Nat Commun 11(1):5749
6. Keyes KM, Westreich D (2019) UK Biobank, big data, and the consequences of non-representativeness. Lancet 393(10178):1297
7. Pirastu N et al (2021) Genetic analyses identify widespread sex-differential participation bias. Nat Genet 53(5):663–671
8. Adams MJ et al (2020) Factors associated with sharing e-mail information and mental health survey participation in large population cohorts. Int J Epidemiol 49(2):410–421
9. Tyrrell J et al (2021) Genetic predictors of participation in optional components of UK Biobank. Nat Commun 12(1):886
10. Stamatakis E et al (2021) Is cohort representativeness Passé? Poststratified associations of lifestyle risk factors with mortality in the UK Biobank. Epidemiology (Cambridge) 32(2):179
11. Yang J et al (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88(1):76–82
12. Yang J et al (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7):565–569
13. Lee JJ et al (2018) Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet 50(8):1112–1121
14. Nagel M et al (2018) Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. Nat Genet 50(7):920–927
15. Howard DM et al (2019) Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci 22(3):343–352
16. Stahl EA et al (2019) Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat Genet 51(5):793–803
17. Pardiñas AF et al (2018) Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet 50(3):381–389
18. Watson HJ et al (2019) Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. Nat Genet 51(8):1207–1214
19. Lloyd-Jones LR et al (2019) Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat Commun 10(1):5086
20. Purcell S et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575
21. Huang JY (2021) Representativeness is not representative: addressing major inferential threats in the UK Biobank and other big data repositories. Epidemiology 32(2):189–193
22. Kendler KS, Heath A, Martin NG (1987) A genetic epidemiologic study of self-report suspiciousness. Compr Psychiatry 28(3):187–196
23. Widiger TA, Oltmanns JR (2017) Neuroticism is a fundamental domain of personality with enormous public health implications. World Psychiatry 16(2):144–145
24. Kaye Walter H, Weltzin T, Hsu LKG (1993) Relationship between anorexia nervosa and obsessive and compulsive behaviors. Psychiatr Ann 23(7):365–373
25. Boraska V et al (2014) A genome-wide association study of anorexia nervosa. Mol Psychiatry 19(10):1085–1094
26. Haworth S et al (2019) Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. Nat Commun 10(1):1–9
27. Zaidi AA, Mathieson I (2020) Demographic history mediates the effect of stratification on polygenic scores. Elife 9:e61548
28. Mitchell BL et al (2021) The Australian genetics of depression study: new risk loci and dissecting heterogeneity between subtypes. Biol Psychiatry 92:227–235