Quantifying between-cohort and between-sex genetic heterogeneity in major depressive disorder.

Maciej Trzaskowski [1], Divya Mehta [1,2], Wouter J. Peyrot [3], David Hawkes [4], Daniel Davies [5] David M. Howard [6], Kathryn E. Kemper [1], Julia Sidorenko [1], Robert Maier [1,7], Stephan Ripke [8,9,10], Manuel Mattheisen [11,12], Bernhard T. Baune [13], Hans J. Grabe [14], Andrew C. Heath [15], Lisa Jones [16], Ian Jones [17], Pamela A.F. Madden [15], Andrew M. McIntosh [6,18], Gerome Breen [19,20], Cathryn M. Lewis [19,21], Anders D. Børglum [12,22], Patrick F. Sullivan [23,24,25], Nicholas G. Martin [26], Kenneth S. Kendler [27], Douglas F. Levinson [28], Naomi R. Wray [1,29], Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium

1 Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia
2 School of Psychology and Counselling, Queensland University of Technology, Brisbane, Australia
3 Department of Psychiatry, Vrije Universiteit Medical Center and GGZ in Geest, Amsterdam, The Netherlands
4 AGRF, The University of Queensland, Brisbane, Queensland, Australia
5 Department of Psychiatry, Behavioural and Clinical Neuroscience Institute and Developmental Psychiatry, Cambridge University, Cambridge, England, United Kingdom
6 Division of Psychiatry, University of Edinburgh, Edinburgh, United Kingdom
7 Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts
8 Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts
9 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts
10 Department of Psychiatry and Psychotherapy, Universitätsmedizin Berlin Campus Charité Mitte, Berlin, Germany
11 Department of Biomedicine, Aarhus University, Aarhus, Denmark
12 iPSYCH, The Lundbeck Foundation Initiative for Integrative Psychiatric Research, Aarhus, Denmark
13 Department of Psychiatry, The University of Melbourne, Melbourne, Victoria, Australia
14 Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany
15 Department of Psychiatry, Washington University in Saint Louis School of Medicine, Saint Louis, Missouri
16 Institute of Health & Society, University of Worcester, Worcester, United Kingdom
17 MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, United Kingdom
18 Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, United Kingdom
19 MRC Social Genetic and Developmental Psychiatry Centre, King's College London, London, United Kingdom
20 NIHR BRC for Mental Health, King's College London, London, United Kingdom
21 Department of Medical and Molecular Genetics, King's College London, London, United Kingdom

22 Department of Biomedicine and iSEQ-Centre for Integrative Sequencing, Aarhus University, Aarhus, Denmark
23 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
24 Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina
25 Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina
26 Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia
27 Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia
28 Psychiatry and Behavioral Sciences, Stanford University, Stanford, California
29 Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia

Major depressive disorder (MDD) is clinically heterogeneous with prevalence rates twice as high in women as in men. There are many possible sources of heterogeneity in MDD most of which are not measured in a sufficiently comparable way across study samples. Here, we assess genetic heterogeneity based on two fundamental measures, between-cohort and between-sex heterogeneity. First, we used genome-wide association study (GWAS) summary statistics to investigate between-cohort genetic heterogeneity using the 29 research cohorts of the Psychiatric Genomics Consortium (PGC; N cases = 16,823, N controls = 25,632) and found that some of the cohort heterogeneity can be attributed to ascertainment differences (such as recruitment of cases from hospital vs. community sources). Second, we evaluated between-sex genetic heterogeneity using GWAS summary statistics from the PGC, Kaiser Permanente GERA, UK Biobank, and the Danish iPSYCH studies but did not find convincing evidence for genetic differences between the sexes. We conclude that there is no evidence that the heterogeneity between MDD data sets and between sexes reflects genetic heterogeneity. Larger sample sizes with detailed phenotypic records and genomic data remain the key to overcome heterogeneity inherent in assessment of MDD.

KEYWORDS
depression, genetic heterogeneity, LD score regression, MDD, sex differences

1 INTRODUCTION

Major depressive disorder (MDD) is a common debilitating disorder with lifetime risk of ~15% (Kessler & Bromet, 2013; Lohoff, 2010). Genetic factors contribute to etiology of MDD with heritability estimated to be ~37% (Kendler, Gatz, Gardner, & Pedersen, 2006; Sullivan, Neale, & Kendler, 2000) of which about one-third is tracked by common-genetic variants (Cross-Disorder Group of the Psychiatric Genomics et al., 2013; Wray et al., 2018). Nongenetic factors also contribute and environmental risk factors include childhood psychological trauma (Chapman et al., 2004; Heim, Newport, Mletzko, Miller, & Nemeroff, 2008; Vythilingam et al., 2002), social isolation (Bruce & Hoff, 1994), and medical conditions, such as cardiovascular disease (Fiedorowicz, 2014; Fraguas et al., 2007; Huffman, Celano, Beach, Motiwala, & Januzzi, 2013). Most complex disorders are considered to be heterogeneous at clinical presentation. For MDD, heterogeneity is inherent in the diagnostic

framework since diagnosis is achieved through different combinations of endorsements of at least five out of nine criteria in the context of depressed mood for most of the day every day for 2 weeks (Diagnostic and Statistical Manual of Mental Disorders [DSM] criteria). Heterogeneity in symptom profiles between individuals reflects not only the symptoms endorsed, but for some criteria (those assessing sleep, weight/appetite, and psychomotor function) the endorsement can reflect either increase or decrease (or both). It is plausible that these clinical differences reflect different biological pathways. The lack of a biological "gold standard" definition in psychiatric illness is well recognized (Kapur, Phillips, & Insel, 2012), and a key question for the field is whether genetic heterogeneity underpins phenotypic heterogeneity (Fanous & Kendler, 2005), and if genome-wide genetic data can support analyses that demonstrate genetic heterogeneity (Han et al., 2016). Here, we assess genetic heterogeneity based on two fundamental measures available to us, between-cohort and between-sex heterogeneity. While nonbiological factors (such as ascertainment strategy) could contribute to both between-cohort and between-sex heterogeneity, evidence for between-sex heterogeneity may reflect, at least in part, biological differences.

Prevalence rates of MDD in women that are double those of men are consistently reported in epidemiological studies, with lifetime risk approximately 0.2 for females and 0.1 for males (Kessler, 2003). Women tend to have younger age of onset, greater comorbidity with panic and other anxiety disorders, whereas men exhibit stronger comorbidity with alcohol dependence or abuse (Schuch, Roest, Nolen, Penninx, & de Jonge, 2014). Attempts to link the epidemiological differences to biological differences have been less consistent. Some twin studies reported significantly higher heritability in females (0.42, 95% CI = 0.36–0.47) than males (0.29, 95% CI = 0.19–0.38), and with genetic correlation significantly different from 1 (r_g~0.60, 95% CI = 0.31–0.99) (Kendler et al., 2006). Other studies failed to find differences between sexes (Fernandez-Pujals et al., 2015). Drawing strong conclusions may be confounded by reporting biases as males are more likely to underreport their symptoms when compared to females (Hunt, Auriemma, & Cashaw, 2003; Thornicroft et al., 2017).

We use genome-wide association study (GWAS) summary statistics data to investigate genetic heterogeneity of MDD. We study between-cohort genetic heterogeneity using data from the 29 independent studies that comprise the Wave 2 PGC-MDD study (PGC29 [Wray et al., 2018]). We also investigate genetic heterogeneity by sex using GWAS summary statistics from PGC29 and three other large data sets. We evaluate between-cohort and between-sex genetic heterogeneity estimates of SNP-heritabilities and genetic correlations. These estimates of genetic parameters, calculated from genome-wide data, provide single statistic summaries of the data. Specifically, differences in SNP-heritability estimates between samples could imply real differences in the relative magnitude of genetic risk effect sizes between samples or could reflect biases due to ascertainment characteristics of the sample. In contrast, an estimate of a genetic correlation less than one may reflect differences in the relative ordering of genetic risk effects between samples. It is possible for SNP-heritabilities to differ between samples but the genetic correlations to be one.

2 MATERIALS AND METHODS

2.1 Between-cohort heterogeneity

We investigate heterogeneity between cohorts from the PGC Working Group for MDD (PGC-MDD) (Major Depressive Disorder Working Group of the Psychiatric et al., 2013), which comprises 29 cohorts (PGC29, 10 from Wave 1 (Major Depressive Disorder Working Group

of the Psychiatric et al., 2013) and 19 from Wave 2 (Wray et al., 2018)), totaling 16,815 cases (68% female) and 25,485 controls (51% female) (Table 1, Supporting Information Table S1). Cohorts represent individual studies in which cases and controls were imputed together to the 1,000 Genomes reference panel (Genomes Project et al., 2010) from a common set of SNPs that had been processed through a common quality control (QC) pipeline (Wray et al., 2018). For the majority of cohorts (but not all), cases and controls were collected by the same research group and were genotyped together on the same genotyping array. All 29 case cohorts passed a structured methodological review by MDD assessment experts (DF Levinson and KS Kendler). Cases were required to meet international consensus criteria (DSM-IV, International Statistical Classification of Diseases [ICD]-9, or ICD-10) (American Psychiatric Association, 1994; World Health Organization, 1978, 1992) for a lifetime diagnosis of MDD established using structured diagnostic instruments from assessments by trained interviewers, clinician-administered checklists, or medical record review. Nonetheless, there were differences in ascertainment across cohorts (Supporting Information Table S1). For example, the RADIANT cohort (rad3) (Lewis et al., 2010) recruited cases of clinically assessed recurrent MDD, which being more severe have lower lifetime risk ~5% (McGuffin, Katz, Watkins, & Rutherford, 1996), compared to community samples such as the QIMR cohorts (qi3c, qi6c, and qio2) assessed by self-report interview and with lifetime risk ~24% (Mosing et al., 2009). To capture heterogeneity due to ascertainment, we coded the 29 cohorts as identified in community, psychiatric outpatient, psychiatric inpatients, or mixed in−/out-patient settings (Supporting Information Table S1).

2.2 Between-sex heterogeneity

We investigate between sex heterogeneity using four large MDD data sets (Table 1). In addition to PGC29, we used the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort (Banda et al., 2015) (where electronic medical records from the Kaiser Permanente healthcare system were used to identify cases as individuals being treated for any psychiatric disorder), the Danish iPSYCH cohort (where national hospital records identified cases as those ever treated clinically for MDD and controls as those who have not), and the volunteer UK Biobank (Bycroft et al., 2018; Lane et al., 2016) (UKB) study. UKB cases were those with either recorded ICD10 codes for MDD (F32, F33) or self-report for seeking treatment for nerves, anxiety or depression; for detailed description of the “broad depression” definition see reference (Howard et al., 2018). Exclusions for both cases and controls were those with recorded schizophrenia, bipolar or mental retardation diagnoses or prescriptions associated with these disorders. Additional exclusions for controls included those with recorded anxiety, phobic or autistic spectrum disorders. In all studies, cases and controls were unrelated. GWAS summary statistics for each cohort used the same methods as for PGC29.

2.3 Statistical methods

We use GWAS summary statistics and linkage disequilibrium (LD) score analysis (LDSC) (Bulik-Sullivan et al., 2015) to estimate the total proportion of variance in liability attributable to SNPs genomewide (i.e., SNP-heritability). Bivariate LDSC was used to estimate the genetic correlation tagged by genome-wide SNPs (rg) between two traits. LDSC has been applied widely to GWAS summary statistics of psychiatric (Anttila et al., 2018) and other disorders (Bulik-Sullivan et al., 2015), and results have been shown to agree well with estimates made from full individual-level genotype and phenotype data using linear mixed

model analysis (e.g., GREML [Yang et al., 2010]), as long as the LD reference sample is drawn from a population that appropriately reflects the samples contributing the GWAS summary statistics for each cohort used the same methods as for PGC29.

2.3 Statistical methods

We use GWAS summary statistics and linkage disequilibrium (LD) score analysis (LDSC) (Bulik-Sullivan et al., 2015) to estimate the total proportion of variance in liability attributable to SNPs genomewide (i.e., SNP-heritability). Bivariate LDSC was used to estimate the genetic correlation tagged by genome-wide SNPs (rg) between two traits. LDSC has been applied widely to GWAS summary statistics of psychiatric (Anttila et al., 2018) and other disorders (Bulik-Sullivan et al., 2015), and results have been shown to agree well with estimates made from full individual-level genotype and phenotype data using linear mixed model analysis (e.g., GREML [Yang et al., 2010]), as long as the LD reference sample is drawn from a population that appropriately reflects the samples contributing the GWAS summary statistics (Yang et al., 2015). A key advantage of LDSC is the minimal computational requirements compared to methods that use individual level data, and the ability to differentiate between genomic inflation due to polygenicity and due to population stratification. Disadvantagesof LDSC are that standard errors (s.e.) of estimates can be (about 50%) higher compared to when estimates are based on full data, particularly for rg estimates (Ni, Moser, Schizophrenia Working Group of the Psychiatric Genomics, Wray, & Lee, 2018).

TABLE 1 Description of GWAS data sets for between-sex heterogeneity analyses

Data set	Cases	Controls	Female cases	Female controls	Male cases	Male controls	Number of Cohorts[a]
PGC29	16,823	25,632	11,438	12,463	5,377	13,022	29[b]
GERA	7,162	38,287	5,152	20,650	2,010	17,637	1
UKB	113,769	208,801	73,292	99,385	40,477	109,426	1
iPSYCH	18,577	17,637	12,690	8,534	5,887	9,103	1
Total	156,331	290,357	102,572	141,032	53,751	149,188	32

[a] Cohort is defined as the cases and controls with genome-wide genotypes imputed from the same set of SNPs that have passed through a common quality control pipeline. Mostly, cohort reflects a case–control sample collected by a PGC principal investigator.
[b] Cohorts ranged in size from 246 to 3,760 cases plus controls.

SNP-heritability is estimated on the observed binary scale h2 SNP − cc, but these estimates depend on the proportion of cases in the sample (P) and so are not easily comparable across cohorts. Hence, for improved interpretability and comparison across studies, h2 SNP − cc is transformed to the liability scale h2 SNP (Lee, Wray, Goddard, & Visscher, 2011) based on normal distribution theory, given an assumed lifetime risk of disease in the population (K):

$$h^2_{SNP} = h^2_{SNP-cc}\frac{(K(1-K))^2}{P(1-P)z^2} \qquad (1)$$

where z is the height of the standard normal density function when truncated at proportion K. However, this transformation assumes that controls are screened. Peyrot, Boomsma, Penninx, and Wray (2016) showed that when the proportion of controls that are unscreened is u, then transformation should be

$$h^2_{SNP} = h^2_{SNP-cc} \frac{(K(1-K))^2}{P(1-P)(1-uK)^2 z^2} \qquad (2)$$

which reduces to Equation 1 when all controls are screened, u = 0. When diseases are uncommon, assuming controls are screened when they are not makes little impact (Peyrot et al., 2016). However, for very common disorders, such as MDD, the difference is not trivial. For example, for K = 0.15, h2 SNP − cc = 0.15, P = 0.5, then h2 SNP= 0.18 when controls are screened and 0.24 when unscreened. The rg estimates are robust to P, K, and u, since these factors contribute to both numerator and denominator of the correlation (which is defined as the estimate of the additive genetic covariance divided by the product of the square root of the SNP-heritabilities for the two traits). Hence rg estimates are robust to ascertainment practices and approximately the same where estimated on the case–control observed scale or liability scales (Bulik-Sullivan, Finucane, et al., 2015). If the same genetic effects contribute to disease risk between sexes or between cohorts then rg is expected to be 1. It was not possible to compare h2 SNP of each PGC29 cohort, because the per-cohort estimates had high s.e. (e.g., a cohort of 500 cases and 500 controls would be expected to produce h2 SNP with standard error of at minimum 0.38 [Visscher et al., 2014]). Instead we estimated the h2 SNP attributed to a cohort by evaluating its contribution to h2 SNP estimates calculated from 500 random samplings of cohorts drawn from the 29 PGC29 cohorts. In each sampling, we randomly selected cohorts until the total sample size was ≥5,000, then used the GWAS summary statistics meta-analyzed (weighted by s.e.) in LDSC to estimate h2 SNP assuming lifetime risk of K = 0.15, and assuming controls are screened (Equation 1). To determine the contribution to the h2 SNP estimate from each cohort we fitted a linear model with estimated h2 SNP as the dependent variable regressed on indicator variables set as 1 if the cohort contributed to the estimate (was included in the random sampling), and 0 otherwise.

3 RESULTS

3.1 Between-cohort heterogeneity within PGC29

We estimated h2 SNP in 500 random samplings of the cohorts from PGC29. From a linear regression of h2 SNP on indicator variables set as 1 if the cohort contributed to the estimate and 0 if it did not, we estimated an h2 SNP effect size deviation per cohort (y-axis Figure 1). Fifteen of the 29 cohorts had h2 SNP deviations different from zero ($p < 0.05/29$). We found that the cohorts nes1 (combined sample of the Netherlands Study of Depression and Anxiety and the Netherlands Twin Registry) (Boomsma et al., 2008; Penninx et al., 2008) and gep3 (GenPod/NEWMEDS) (Lewis et al., 2011) contributed most to variation in estimates of h2 SNP, and explain 0.14 and 0.16, respectively, of the variance in h2 SNP estimates across the 500 samplings. Samplings that included cohort nes1 had the highest average estimates of h2 SNP, while samplings including gep3 had the lowest average estimates. These differences are in line with expectations based on screening strategies for controls (Supporting Information Table S1). The nes1 cohort used super-screened controls (Boomsma et al., 2008), such that controls never scored higher than 0.65 on a general factor score for anxious depression (mean = 0, SD = 0.7) derived from a combined measure of neuroticism, anxiety, and depressive symptoms assessed via longitudinal questionnaires over 15 years. In contrast, the gep3 cohort was a case-only research cohort which was

matched to independently collected and genotyped controls (hence particularly stringent QC is needed to combine the genotype data of the contributing cases and controls).
In fact, gep3 is one of seven cohorts for which controls were unscreened for MDD (Figure 1), but only one other cohort used independently genotyped controls (STAR*D, coded as stm2); together the seven cohorts have lower mean beta-values, but not significantly so (p = 0.055). The trend in these results might be explained by recognizing that SNPheritability is first estimated on the observed binary case–control scale h2 SNP − cc and then transformed to the liability scale h2 SNP. Indeed, we find that increasing sample prevalence (P in Equation 1) is significantly associated with the estimated h2 SNP (p = 0.00057), but not sex ratio (p = 0.72). The application of the standard transformation (Equation 1), as we have done, assumes screened controls and could generate an underestimate of the SNP-heritability if controls were in fact unscreened. Similarly, super-screening of controls could generate an over-estimate of the true h2 SNP. Hence, we expect that the standard transformation would generate an overestimate for the nes1 cohort (super-screened controls) and an underestimate for cohorts with unscreened controls, consistent with our results. Next, we investigated if h2 SNP estimates differed based on the research protocol to ascertain cases. For the same proportion of cases and controls in the GWAS sample, we would expect the h2 SNP − cc to be higher for a clinically ascertained cohort than a community ascertained cohort, further we would expect the transformation based on K = 0.15 (Equation 1) to overestimate h2 SNP when the true K is lower (clinical cohort) and underestimate h2 SNP when the true K is higher (community cohort). There is evidence to support this hypothesis (Figure 1). We found significant difference between the mean estimates of community (−0.027, s.e. 0.007) vs noncommunity cohorts (−0.08 s.e. 0.006) (with noncommunity comprising the three in- and out-patient categories), using a one-sided, two-sample t test assuming unequal variance (p = 0.028) (Supporting Information Table S4). The difference became more significant (p = 0.015) when the cohorts we had a priori reason to exclude, namely nes1 and gep3, based on discussions above were removed.

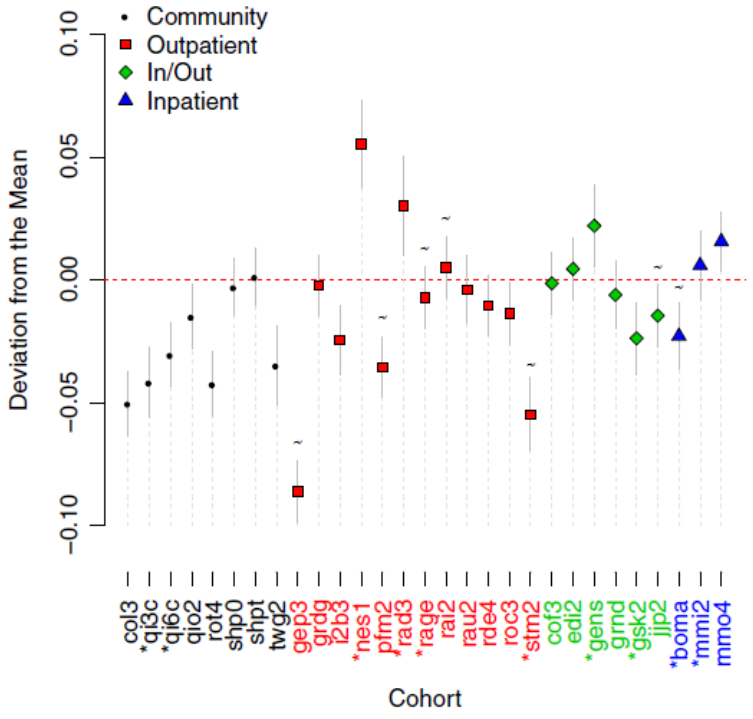


FIGURE 1 Cohort deviation estimates from the linear regression of h2 SNP estimates (from each of the 500 samplings of cohorts) on cohort indicator variables set at 1 if the cohort was included in the sampling that generated the h2 SNP and 0 otherwise. In each sampling, cohorts were selected at random until the total case/control sample size exceeded 5,000. Cohort GWAS results were meta-analyzed and these results passed into LDscore. h2 SNP was estimated using Equation 1 transformation (K = 0.15) which assumes controls are screened. h2 SNP estimates of samplings were highest, on average, when cohort nes1 was included and lowest, on average, when cohort gep3 was included. Wave 1 cohorts have an asterisk by their name and cohorts that have unscreened controls are marked by a tilde. Continuous lines around data-points are 95% confidence Intervals. For explanation of cohort names see Supporting Information Table S1

TABLE 2 Estimates of h^2_{SNP} from LDSC applied to sex-specific GWAS summary statistics

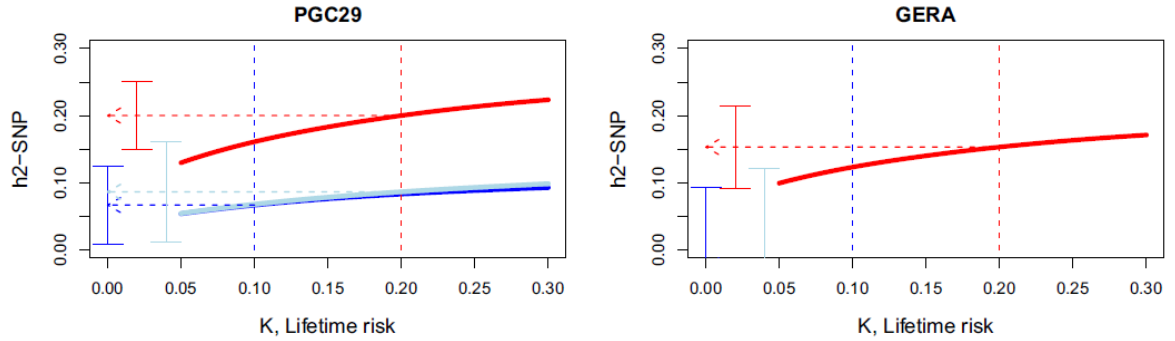
	Female (s.e.)		Males v1 (s.e.)		Males v2 (s.e.)		*p*-value v1	*p*-value v2
K	0.2		0.1		0.2			
u	0		0		0.1			
PGC29	0.20	(0.03)	0.07	(0.04)	0.09	(0.05)	0.61	0.68
GERA	0.15	(0.04)	−0.02	(0.05)	−0.03	(0.07)	0.55	0.57
UKB	0.10	(0.01)	0.07	(0.01)	0.10	(0.01)	0.77	0.94
iPSYCH	0.23	(0.03)	0.15	(0.04)	0.20	(0.05)	0.77	0.91
Meta-4	0.11	(0.005)	0.07	(0.006)	0.10	(0.007)	1.6×10^{-6}	0.10
Meta-6	0.10	(0.005)	0.07	(0.006)	0.10	(0.008)	1.2×10^{-3}	0.60
Meta-10	0.11	(0.004)	0.07	(0.004)	0.10	(0.005)	1.1×10^{-8}	0.12
GWAS-Meta	0.08	(0.004)	0.06	(0.005)	0.08	(0.006)	6.6×10^{-4}	0.64

Note. h^2_{SNP} estimates are presented on the liability scale achieved through transformation of the LDSC h^2_{SNP-cc} estimate accounting for the case prevalence in the sample (P), the lifetime risk (*K*) of the disorder, and the proportion of cases in the control sample (*u*), Equation 2. Meta-4: meta-analysis of the h^2_{SNP} estimates for the 4 data sets (PGC29, GERA, UKB, iPSYCH). Meta-6: meta-analysis of the 6 h^2_{SNP} estimates derived from the genetic covariance estimates from bivariate LDSC between the 6 possible same-sex data set pairwise combinations. Meta-10: meta-analysis based on all h^2_{SNP} estimates contributing to Meta-4 and Meta-6. GWAS-Meta: h^2_{SNP} estimated from the GWAS summary statistics of the 4 data sets. Versions v1 and v2 differ by *K* and *u* values; v2 hypothesis is that the lifetime risk of MDD is the same in men and women but that more cases go unreported in men, and hence cases could be included in a screened control set.

3.2 Between-sex heterogeneity

Using the four large data sets (Table 1) we investigate sex-specific heterogeneity. We used bivariate LDSC to estimate the rg between all pairs of the two sexes by four data sets, but the standard errors were high (Supporting Information Table S2). rg involving the GERA_M data set were not estimable, because of the negative/zero of h2 SNP used in the denominator of the rg estimate. The between-sex rg estimated from the meta-analysis of the GWAS summary statistics of the 4 data sets was 0.86 (s.e. 0.04; pH0:rg=1= 3.0 × 10-4), and the meta-analysis of 12 male–female rg estimates was 0.76 (s.e. 0.03; pH0:rg = 1 = 8.9 × 10−16). At face value these results imply genetic factors are only partially shared between the sexes. However, this interpretation should be considered with caution when benchmarked by the meta-analysis of 6 female–female rg estimates of 0.72 (s.e. 0.04; pH0:rg = 1 = 4.9 × 10−11) and the metaanalysis of 3 male–male rg estimates of 0.71 (s.e. 0.11; pH0:rg = 1 = 0.11) Hence, the between-sex estimate of rg being significantly different from zero likely reflects the general heterogeneity between the data sets rather than being sex-specific. Next, we investigated sex-specific estimates of h2 SNP using LDSC (Table 2, Supporting Information Table S3) to determine if there is evidence for a greater genetic contribution to MDD risk in females then males. We have power to detect differences of the

order of 2*(s.e. of male estimate + s.e. of female estimate). Initially, in the transformation of the h2 SNP − cc estimate to the liability scale (Equation 1) we assumed K = 0.20 for females and K = 0.10 for males (Table 2), consistent with literature reports that MDD is twice as common in females as males (Weissman, Leaf, Holzer, Myers, & Tischler, 1984). The h2 SNP estimates were smaller for males (range −0.02 to 0.15) than for females (range 0.10 to 0.23), but given the magnitude of the standard errors, none of the h2 SNP sex differences were significantly different for any individual data set. However, meta-analysis of the estimates of the four data sets did lead to estimates that were significantly different (Meta-4 in Table 2; 0.07 in males vs. 0.11 in females, p = 1.6 × 10−6). In addition, h2 SNP estimated from the meta-analyzed GWAS results of the four data sets also showed significant difference between males and females (0.06 vs 0.08, p = 6.6 × 10−4; Table 2 GWAS-Meta). We also meta-analyzed the six h2 SNP values estimated from the genetic covariance between pairs of same-sex data sets in bivariate LDSC analysis. As the traits are (presumed to be) the same, the genetic covariance is also an estimate of genetic variance (Supporting Information Table S3; Table 2 Meta-6). This again showed lower mean estimates for males with a significant difference between the sexes (0.07 in males vs 0.10 in females, p = 0.0012). For completeness, a metaanalysis from all 10 of the estimates is provided (Table 2 Meta-10); this uses the same data sets as the GWAS-Meta, but the latter uses all the information jointly rather than pairwise. Before drawing strong conclusions from these results, it is important to recognize that the estimates of h2SNP depend on the choice of the lifetime risk estimates (K in Equations 1 and 2) (Figure 2). The point estimates are more similar if the same lifetime risk is assumed between the sexes, but it is difficult to justify such an assumption, because it is not, at face value, supported by epidemiological data. However, since depression maybe underreported in males (Martin, Neighbors, & Griffith, 2013; Thornicroft et al., 2017), for illustration purposes we could assume the true lifetime risk of MDD is the same between the sexes (K = 0.20), but that through underreporting the controls are contaminated by 0.10 of cases (Equation 2, u = 0.1). Under these assumptions, the h2 SNP estimates are not significantly different between the sexes for any data set (Figure 2, Table 2). For completeness, we also estimated X-chromosome SNPheritability from the meta-analyzed cohorts for males and females separately. However, the standard errors of the estimates were large relative to the h2 SNP estimates (h2 SNP males=0.0025 (s.e. = 0.06); h2 SNP females=0.0005 (s.e. = 0.03), which meant estimation of the rg between them was not meaningful.



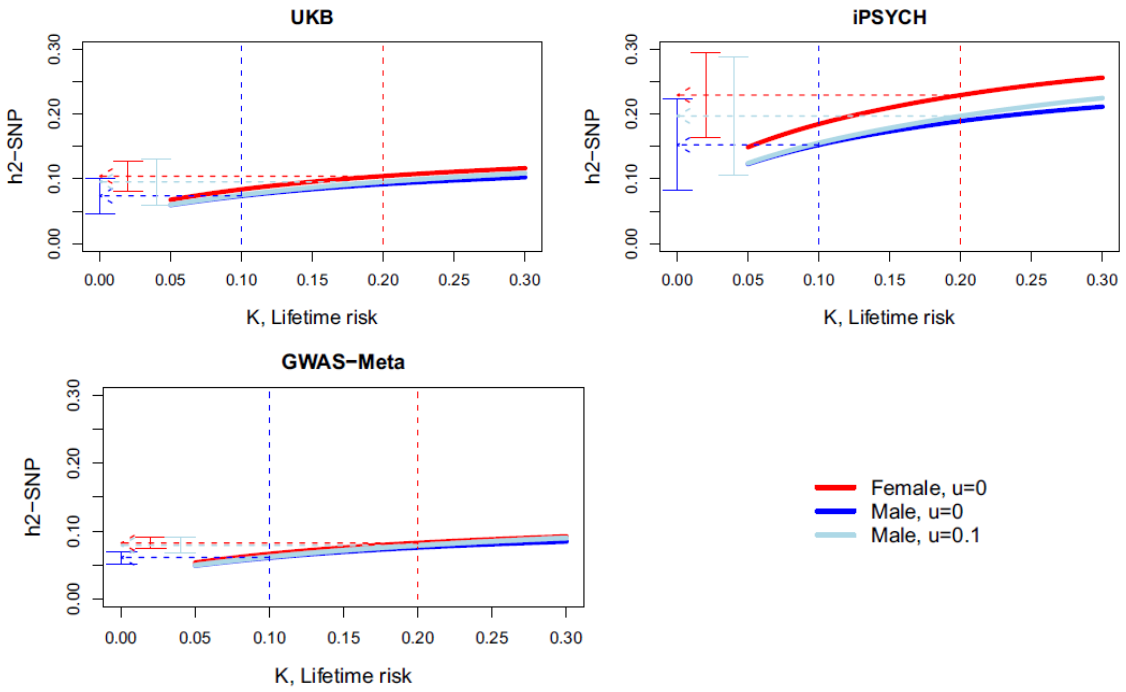


FIGURE 2 Impact of choice of lifetime risk on estimate of h^2_{SNP}. The graphs shows h^2_{SNP} on the liability scale from Equation 2, *u* (proportion of controls that are unrecognized cases). The blue/red dashed lines are positioned at the lifetime risk for males/females. The flat ended bars show the 95% confidence intervals of the h^2_{SNP} estimates at the chosen lifetime risk [Color figure can be viewed at wileyonlinelibrary.com]

4 DISCUSSION

Heterogeneity in MDD is often discussed, but hard to investigate. In a novel set of analyses, we explored the heterogeneity of MDD using genetic data. The first set of analyses contrasted 29 PGC cohorts, by estimating their average contribution to estimates of h2 SNP from repeated random samplings of cohorts selected into GWAS meta-analyses. While we found notable differences between cohorts in the h2 SNP contribution estimates (Figure 1), these differences could be explained, at least partly, via knowledge of cohort ascertainment practices: higher contributions for cohorts ascertained in clinical compared to community settings (Figure 1, p = 0.028), higher contribution from a sample known to use super-screened controls (nes1), and a trend toward lower contributions from samples that used unscreened controls. One conclusion is that known cohort information about case ascertainment status could be included usefully in analysis methods to increase power. A framework for such an analysis has been proposed (Zaitlen et al., 2012), but in practice the necessary parameters relating to cohort specific risks are usually unknown.In the seven samples contributing to the published PGC metaanalysis (PGC29, GERA, iPSYCH, UK Biobank, deCode, Generation Scotland, 23andMe) (Wray et al., 2018), h2 SNP estimates ranged from 0.09 to 0.25 and the weighted mean rg for all pairwise combinations was 0.76 (s.e. = 0.03), which is significantly different from one. The cohorts had different recruitment strategies with ascertainment ranging from self-report to national hospital records. Moreover, even within the Wave 1 PGC-MDD research cohorts endorsement proportions of the nine DSMIV criteria showed considerable heterogeneity including between cohorts that had similar clinical ascertainment strategies (Major Depressive Disorder Working Group of the Psychiatric et al., 2013). For example, endorsement rates of 56%, 27%, and 10% were recorded for the criterion symptom 4b, hypersomnia nearly every day, for different early onset (<30 years) recurrent MDD samples

(Major Depressive Disorder Working Group of the Psychiatric et al., 2013). Despite the heterogeneity, out-of-sample prediction demonstrated that the self-reported 23andMe GWAS results explained variance in clinically ascertained cohorts with high significance (Wray et al., 2018). Sample size remains the driving force for genetic discovery in MDD. Ideally, larger sample sizes should be accompanied by collection of detailed, consistent, and longitudinal phenotypic data to enable more precise case and control definitions.

We also investigated between-sex genetic heterogeneity. Our sex-specific analyses found significantly smaller h2 SNP for males than females, a trend replicated in all four data sets, and hence was highly significant in the meta-analysis of the four cohort estimates (Table 2 , male v1). However, we recognized that the comparisons of h2 SNP between the sexes depended on the choice of their respective lifetime risks (Figure 2). For baseline analyses we used lifetime risk estimates of K = 0.20 for females and K = 0.10 for males, consistent with a 2:1 risk for females versus males (Weissman et al., 1984), with higher K values generating higher h2 SNP estimates. One explanation for a lower lifetime risk for males could be higher rates of underreporting (Martin et al., 2013; Thornicroft et al., 2017). We calculated h2 SNP in males assuming the same lifetime risk as females, but with incomplete screening of controls. Such a hypothetical scenario generated similar estimates of h2 SNP between the sexes (Figure 2, Table 2).

In summary, our analyses demonstrate between-cohort genetic heterogeneity, but this can be explained, at least in part, by known factors such as case/control ascertainment. Investigation of between sex heterogeneity provided no convincing evidence to support genetic differences between the sexes. A robust conclusion is simply that large sample sizes will overcome sample heterogeneity as demonstrated in the latest major depression GWAS meta-analyses (Howard et al., 2018; Wray et al., 2018). Based on differences in lifetime disease risk and differences in heritability, while assuming a similar number of contributing risk loci, we previously estimated that sample sizes for GWAS need to be five times bigger for MDD than for schizophrenia (SCZ) (Wray et al., 2012). On the one hand, heterogeneity between samples may push this estimate higher. On the other hand, the heterogeneity may already account for the higher prevalence and lower heritability. The PGC GWAS meta-analysis for MDD/major depression based on a total effective sample size of 389,083 (Wray et al., 2018) identified 44 independent significant loci. This compares to 145 independent loci for SCZ from a total effective sample size of 99,863 (Pardiñas et al., 2018), hence requiring >12 times the sample size for major depression compared to SCZ per genome-wide significant locus. However, the relationship between sample size and variant discovery is not linear (Wray et al., 2018) and so observing the sample size ratios for discovery will be of interest as sample sizes increase. Very large MDD case–control samples will allow novel methods to be applied to assess evidence for genetic subsets, and will allow more robust conclusions to be drawn about between sex differences. Larger data sets are likely to lead to the development of new methods to assess genetic heterogeneity (Han et al., 2016). There is a growing interest in machine learning methods (Libbrecht & Noble, 2015) as a strategy to identify phenotypically relevant genetic subsets, but cohort heterogeneity must diminish their utility, making large electronic health or biobank samples collected and genotyped in a uniform way of most value.