Genetic stratification of depression in UK Biobank suggests a subgroup linked to age of natural menopause

David M. Howard*[1,2], Lasse Folkersen[3,4], Jonathan R. I. Coleman[1,5], Mark J. Adams[2], Kylie Glanville[1], Thomas Werge[3,4,6,7], Saskia P. Hagenaars[1,5], Buhm Han[8], David Porteous[9,10], Archie Campbell[9,11], Toni-Kim Clarke[2], Gerome Breen[1,5], Patrick F. Sullivan[12,13,14], Naomi R. Wray[15], Cathryn M. Lewis[1,5], and Andrew M. McIntosh[2,16]

Affiliations:
[1] Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, UK
[2] Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK
[3] Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Copenhagen, Denmark
[4] Institute of Biological Psychiatry, Mental Health Services Capital Region of Denmark, Copenhagen, Denmark
[5] NIHR Maudsley Biomedical Research Centre, South London and Maudsley NHS Trust, London, UK
[6] Department of Clinical Sciences, University of Copenhagen, Copenhagen, Denmark
[7] Lundbeck Foundation's Center for GeoGenetics, GLOBE Institute, University of Copenhagen, Denmark
[8] Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, Republic of Korea
[9] Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, UK
[10] Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK
[11] Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK
[12] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[13] Department of Genetics, University of North Carolina, Chapel Hill, NC, USA
[14] Department of Psychiatry, University of North Carolina, Chapel Hill, NC, USA
[15] Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia
[16] Department of Psychology, University of Edinburgh, Edinburgh, UK

* Corresponding author: David M. Howard
Social, Genetic and Developmental Psychiatry Centre,
Institute of Psychiatry, Psychology & Neuroscience,
King's College London, UK
+44 (0)20 7848 5433
E-mail: David.Howard@kcl.ac.uk

Abstract

Depression is a common and clinically heterogeneous mental health disorder that is frequently comorbid with other diseases and conditions. Stratification of depression may align sub-diagnoses more closely with their underling aetiology and provide more tractable targets for research and effective treatment. In the current study, we investigated whether genetic data could be used to identify subgroups within people with depression using the UK Biobank. Examination of cross-locus correlations was used to test for evidence of subgroups by examining whether there was clustering of independent genetic variants associated with eleven other complex traits and disorders in people with depression. We found evidence of a subgroup within depression using age of natural menopause variants ($P = 1.69 \times 10^{-3}$) and this effect remained significant in females ($P = 1.18 \times 10^{-3}$), but not males ($P = 0.186$). However, no evidence for this subgroup ($P > 0.05$) was found in Generation Scotland, iPSYCH, a UK Biobank replication cohort or the GERA cohort. In the UK Biobank, having depression was also associated with a later age of menopause (beta = 0.34, standard error = 0.06, $P = 9.92 \times 10^{-8}$). A potential age of natural menopause subgroup within depression and the association between depression and a later age of menopause suggests that they partially share a developmental pathway.

Introduction

Depression is a common mental health disorder characterised by persistent feelings of sadness or a loss of interest in day-to-day activities lasting for at least a two-week period. These feelings can be accompanied by tiredness, changes in appetite, changes in sleep patterns, reduced concentration, feelings of worthlessness or hopelessness, and thoughts of self-harm or suicide. Zimmerman et al. [1] found that there were 170 different symptom profiles amongst 1566 participants diagnosed with major depressive disorder from the Rhode Island MIDAS project. This variety of different symptom profiles suggest that depression is highly heterogeneous [2]. Depression is also comorbid with many diseases including cancer [3], cardiovascular disease [4] and other psychiatric illnesses [5].

Stratification of depression, to address heterogeneity and comorbidity, may aid in providing valuable aetiological insights and improve treatment efficacy.

Studies aimed at stratifying depression have examined differences between melancholic and atypical depression [6], differences between the sexes and recurrence of the disorder [7] and used data from other traits, such as neuroticism [8] and social contact [9] to stratify depression. Twin-based studies [10] and genome-wide association studies [11, 12] have shown depression to be heritable and genetically correlated with a number of other traits and disorders. This shared genetic component could be due to pleiotropic variants shared across all individuals but could also be as a result of a subgroup for the other trait within depression cases. For example, there is a genetic correlation of -0.11 (standard error = 0.03) between depression and age of natural menopause [13]. If this genetic correlation was due to pleiotropy, then several of the age of menopause variants would be carried by most depression cases. However, if this correlation was due to a subgroup, then a greater proportion of the age of menopause variants would only be carried by individuals in this subgroup. A subgroup could arise where there is a causal association, a shared molecular pathway, a misclassification between the traits, or an ascertainment bias in the diagnosis of depression.

For the current study, BUHMBOX (Breaking Up Heterogeneous Mixture Based On cross(X)-locus correlations) [14] was used to determine whether there was evidence of a subgroup within depression that was genetically more similar to other traits. BUHMBOX uses variants associated with a non-depression trait to calculate weighted pairwise correlations of risk allele dosages within depression cases and controls, adjusted for effect size and allele frequency. Where there is a subgroup amongst depression cases that carry a greater proportion of the risk alleles for the non-depression trait, there will be consistent positive pairwise correlations between those variants (Figure 1). BUHMBOX then calculates a *P*-value based on the likelihood of the observed pairwise correlations between variants.

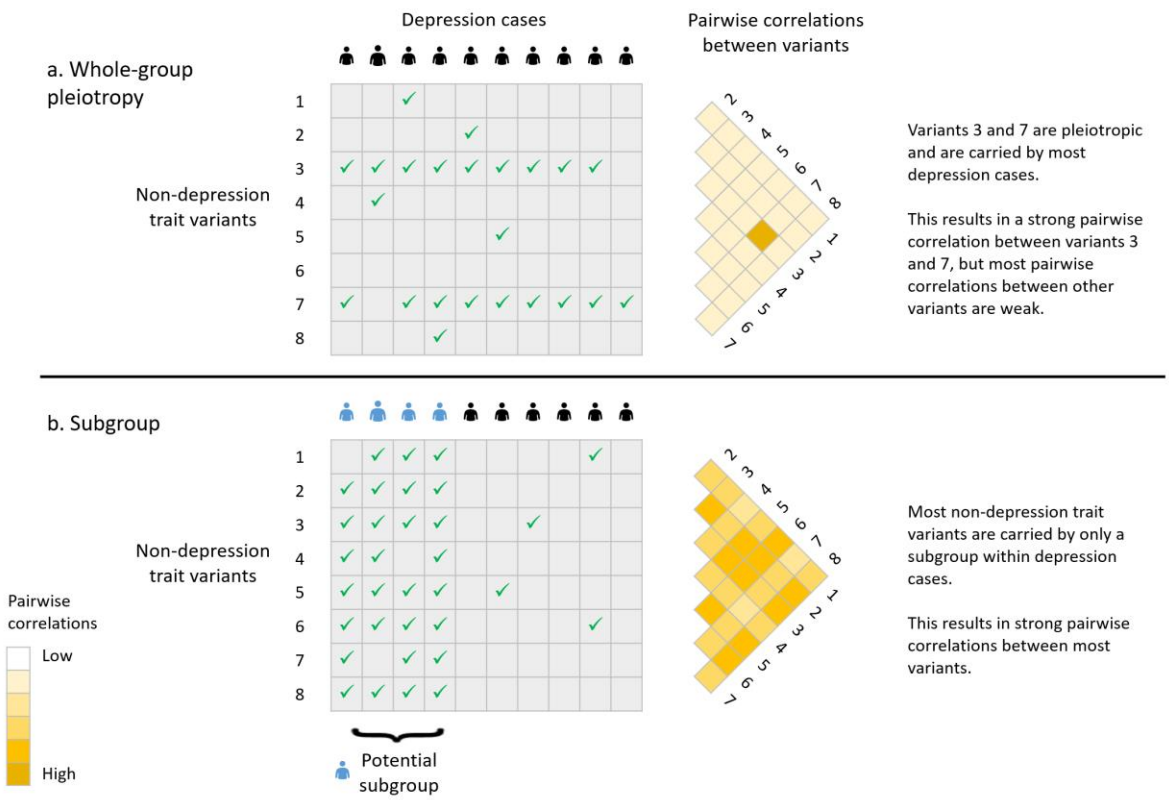

Figure 1. Pairwise correlations between variants for (a) whole-group pleiotropy, where most depression cases carry a few variants associated with a non-depression trait and (b) a subgroup within depression cases ( ), where just the subgroup carry many of the non-depression trait variants. A tick indicates a depression case individual is a carrier of that non-depression variant.

Two definitions of depression were assessed in the UK Biobank [15], one based on the Composite International Diagnostic Interview Short Form (CIDI-SF) [16] and the other based on a broader help-seeking definition (broad depression) [12]. Since many traits are genetically correlated with depression [13], a power calculation was performed to determine traits with sufficient power to detect a subgroup. Power is determined by the number of depression cases, the size of any subgroup within depression cases, the number of associated variants tested from the non-depression trait and the effect sizes of these variants. We tested adequately-powered traits for evidence of a subgroup in depression cases using BUHMBOX v0.38 [14]. Replication of traits forming a subgroup in depression were sought in Generation Scotland: Scottish Family Health Study (GS:SFHS), The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), a UK Biobank replication cohort, and the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. UK Biobank and

106 GS:SFHS were used to investigate phenotypic associations between depression and traits forming a
107 subgroup.

108 **Materials and Methods**

109 UK Biobank discovery cohort

110 The UK Biobank is a population-based cohort of 501,726 individuals with imputed genome-wide data
111 for 93,095,623 autosomal genetic variants [15]. A genetically homogeneous sample of 462,065
112 individuals was identified using the first two principal components from a 4-means clustering
113 approach. A total of 131,790 individuals were identified as being related up to the third degree (kinship
114 coefficients > 0.044) using the KING toolset [17] and were removed from the sample. For these related
115 individuals a genomic relationship matrix was calculated to enable the identification of one individual
116 from each related group that could be reinstated. This allowed the reintroduction of 55,745 individuals
117 providing an unrelated sample of 386,020 individuals.

118 UK Biobank depression phenotypes

119 Two depression phenotypes were assessed for evidence of subgroups in UK Biobank. For the UK
120 Biobank discovery cohort, both phenotypes were restricted to only those individuals that had
121 completed the online mental health questionnaire (n = 109,049). The first phenotype analysed was
122 based on the Composite International Diagnostic Interview Short Form (CIDI-SF) [18] as used by Davis
123 et al. [16] to provide a lifetime instance measure of depression in the UK Biobank. Davis et al. [16]
124 provide a more in-depth description of this CIDI-SF phenotype, but in summary cases were defined as
125 having:

126 - at least one core symptom of depression (persistent sadness (Data-Field: 20446) or a loss of
127 interest (Data-Field: 20441)) for most or all days over a two-week period which were present
128 "most of the day" or "all of the day".
129 - plus at least another four non-core depressive symptoms with some or a lot of impairment
130 experienced during the worst two-week period of depression or low mood.

131 The non-core depressive symptoms that were included in this assessment of the worst episode of
132 depression were: Feelings of tiredness (Data-Field: 20449), Weight change (Data-Field: 20536), Did
133 your sleep change? (Data-Field: 20532), Difficulty concentrating (Data-Field: 20435), Feelings of
134 worthlessness (Data-Field: 20450), and Thoughts of death (Data-Field: 20437). Cases that self-
135 reported another mood disorder were excluded. Controls were determined by not having at least one
136 core symptom of depression or not endorsing at least another four non-core depressive symptoms if
137 at least one core symptom was endorsed. This provided a total of 25,721 CIDI-SF cases and 61,894
138 controls.

139 A second depression phenotype within the UK Biobank discovery cohort was also examined using the
140 broad depression definition from Howard et al. [12] with detailed information provided in that paper.
141 In summary, cases had sought help for nerves, anxiety, tension or depression from either a general
142 practitioner or a psychiatrist (Data-Field: 2090 and Data-Field: 2100), whereas controls had not. Cases
143 were supplemented with an additional 132 individuals identified as having a primary or secondary
144 International Classification of Diseases (ICD)-10 diagnosis of a depressive mood disorder from linked
145 hospital admission records (Data-Field: 41202 and Data-Field: 41204). Participants identified with
146 bipolar disorder, schizophrenia or personality disorder and those reporting a prescription for an
147 antipsychotic medication were removed. This provided a total of 36,790 broad depression cases and
148 70,304 controls. The phenotypic correlation between the CIDI-SF depression phenotype and the broad
149 depression phenotype was 0.61 with the number of cases and controls shared across the two
150 definitions shown in Supplementary Table 1.

151 Traits examined as subgroups within depression

152 We selected traits genetically correlated with depression (false discovery rate corrected, $q < 0.01$) in
153 Howard et al. [13] to test as subgroups within depression, which included anthropomorphic,
154 autoimmune, life course, cardiovascular and other psychiatric traits. For each trait, there was a

155 requirement that publicly available summary statistics were available and that the UK Biobank was
156 not included in that study due to potential confounding effects (Supplementary Table 2).

157 The BUHMBOX power calculation test v0.1 [14] was used to determine whether there was sufficient
158 power to detect a subgroup for each depression correlated trait and to identify the optimum variant
159 selection criterion ($P < 5 \times 10^{-8}$, $P < 10^{-6}$ or $P < 10^{-4}$). The power calculation was conducted using the
160 CIDI-SF depression phenotype and then using the broad depression phenotype. Variants from the
161 summary statistics for each non-depression trait were examined in the UK Biobank discovery cohort.
162 Variants that had a call rate less than 0.99, were out of Hardy-Weinberg equilibrium ($P < 10^{-10}$), had a
163 hard call threshold less than 0.25, or had a minor allele frequency less than 0.005 were excluded.
164 BUHMBOX requires that all variants are available for all individuals and therefore individuals with a
165 call rate less than 1 were removed. To identify independently segregating variants, clumping was
166 conducted in PLINK v1.90b4 [19] using an r^2 value of 0.01 across a 3Mb window in either CIDI-SF or
167 broad depression control individuals, respectively.

168 For the power analysis the approach used in Han et al. [14] was followed, with 1000 simulated
169 iterations run for each trait, the proportion of individuals in the subgroup was set to 0.2 and a nominal
170 subgroup *P*-value of 0.05 was used. Power analyses were used to identify the optimum variant
171 selection criterion that provided the greatest power for each non-depression trait. Where power was
172 the same across variant selection criteria, the strictest variant selection criterion was selected as the
173 optimum. Variants with $P < 10^{-4}$ were not publicly available for Squamous Cell Lung Cancer or Lung
174 Cancer and so $P < 10^{-5}$ was used instead. Only those traits that had a power > 0.8 (using the optimum
175 variant selection criterion) were selected to be tested for evidence of a subgroup within depression.

176 Testing for subgroups within depression

177 For the traits that had power > 0.8, variants meeting the optimum variant selection criterion were
178 extracted from the UK Biobank discovery cohort. The same quality control thresholds and method to
179 identify independently segregating variants as used as previously in the power analysis were applied.

180 BUHMBOX v0.38 [14] was used to examine shared risk alleles for each non-depression trait within
181 CIDI-SF depression and broad depression. BUHMBOX uses the positive correlations between risk allele
182 dosages in cases to determine whether any sharing of risk alleles is driven by all individuals (whole-
183 group pleiotropy) or by a subset of individuals (Figure 1). The likelihood of observing such positive
184 correlations are used to determine the subgroup *P*-values. The BUHMBOX software and manual are
185 freely downloadable from http://software.broadinstitute.org/mpg/buhmbox/.

186 Sex, age, genotyping array and the first 20 principal components were fitted as covariates in the
187 subgroup analysis. Bonferroni correction was used to account for the multiple testing of non-
188 depression traits, with *P*-values $< 5 \times 10^{-3}$ (0.05/10) or $< 4.5 \times 10^{-3}$ (0.05/11) deemed significant for
189 CIDI-SF or broad depression, respectively. No multiple testing correction was applied for the two
190 depression definitions analysed.

191 BUHMBOX calculates and outputs polygenic risk scores for each individual based on the summary
192 statistics provided. If a subgroup for a trait exists, as in Figure 1b, then potentially this subgroup would
193 carry a greater number of these variants compared to the non-subgroup depression cases and
194 therefore a binomial distribution would exist within the polygenic risk scores of cases. To examine
195 whether the standardised distributions of polygenic risk scores for non-depression traits in depression
196 cases and controls could be explained by two univariate normal distributions the mix2normal function
197 from the VGAM package [20] in R v3.5.2 was used. The use of polygenic risk scores provides additional
198 supporting evidence of a subgroup and provides an estimation of the size of any subgroup.

199 Replication of significant subgroups within depression

200 Traits that showed significant evidence of forming a subgroup for depression in the UK Biobank
201 discovery cohort were re-examined in independent cohorts: Generation Scotland: Scottish Family
202 Health Study (GS:SFHS), The Lundbeck Foundation Initiative for Integrative Psychiatric Research
203 (iPSYCH), a UK Biobank replication cohort, and the Genetic Epidemiology Research on Adult Health
204 and Aging (GERA) cohort. In each of the replication cohorts, individuals were removed if they had a

205 variant call rate less than 1 and variants were removed if they had a call rate less than 0.99, were out
206 of Hardy-Weinberg equilibrium ($P < 10^{-10}$) or had a minor allele frequency less than 0.005.

207 The family and population-based GS:SFHS cohort [21] consisted of 23,960 individuals, of whom 20,195
208 were genotyped and subsequently imputed [22] providing a total of 8,633,288 variants for 20,032
209 individuals (11,085 females and 8,947 males). An unrelated subset was created using GCTA v1.22 [23]
210 ensuring that no two individuals shared a genomic relatedness of ≥ 0.025. Individuals were removed
211 if they were identified as population outliers [24] or had participated in UK Biobank (using a checksum-
212 based approach [25]). Sex, age and the first 20 principal components were fitted as covariates in the
213 subgroup tests. A diagnosis of major depressive disorder (MDD) was made using two initial screening
214 questions and the Structured Clinical Interview for the Diagnostic and Statistical Manual of Mental
215 Disorders [26] and has been described previously in Fernandez-Pujals et al. [27]. Using record linkage
216 to the Scottish Morbidity Record, we removed 1,072 controls who had attended at least one
217 psychiatry outpatient clinic. Using the psychiatric inpatient records, we identified 47 MDD cases who
218 were also diagnosed with bipolar disorder or schizophrenia and these individuals were excluded. This
219 provided a total of 975 MDD cases and 5,971 controls. These participants provided prior consent for
220 their anonymised data to be linked to medical records.

221 iPSYCH is a case-control sample with genotyping data collected for 77,639 individuals after quality
222 control. The iPSYCH sample was phased using SHAPEIT3 [28] and imputed using Impute2 [29] using
223 the 1000 genomes phase 3 data [30]. An unrelated subsample was identified using the KING toolset
224 [17] with second degree relatives or closer excluded. Sex, age, genotyping array and the first 20
225 principal components were fitted as covariates in the subgroup tests. Depression status was
226 ascertained from in- and out- patient hospital records with controls screened to ensure they had no
227 other psychiatric disorders. This provided a total of 19,644 cases and 21,295 controls. Further detailed
228 information on the iPSYCH sample is available in Pedersen et al. [31].

229 The UK Biobank discovery sample consisted of only those individuals that completed the mental health
230 questionnaire. Therefore, the individuals that did not complete the questionnaire were used as an
231 independent replication cohort. For these individuals only the broad depression definition could be
232 assessed and applying the same quality control criteria used in the UK Biobank discovery cohort
233 resulted in 71,282 broad depression cases and 128,303 controls.

234 The GERA cohort is a genotyped subsample of 78,419 participants from the Kaiser Permanente
235 Medical Care Plan, Northern California Region (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-
236 bin/study.cgi?study_id=phs000674.v3.p3). GERA was genotyped using custom designed Affymetrix
237 Axiom arrays [32] before being phased with SHAPEIT v2.5 [33] and imputed with IMPUTE2 v2.3.1 [29]
238 using the 1000 Genomes Project [30] as the reference panel. GERA is a mixed ancestry cohort [34] and
239 in the current analysis only those individuals with European ancestry were examined with related
240 individuals up to the third degree removed. Sex, age and the first 10 principal components were fitted
241 as covariates in the subgroup tests. MDD status was ascertained using ICD-9 coding from in- and out-
242 patient hospital records which identified 4,912 MDD cases and 33,902 controls.

243 During the replication analysis, the BUHMBOX power calculation test v0.1 [14] was applied assuming
244 the same estimated proportion of individuals in the subgroup and the same variant selection criterion
245 as in the UK Biobank discovery cohort, and a nominal subgroup *P*-value of 0.05. BUHMBOX v0.38 [14]
246 was then used to examine whether there was evidence of a subgroup within depression cases in the
247 GS:SFHS, iPSYCH, UK Biobank replication and GERA cohorts. The power and subgroup analyses were
248 run using all individuals and then, as age of natural menopause trait generated a significant result in
249 the subgroup analysis the analyses, were run using females only.

250 Phenotypic examination of significant subgroups within depression

251 The age of natural menopause trait generated a significant result in the subgroup analysis in the UK
252 Biobank discovery cohort and we therefore examined whether those with depression had a later or
253 earlier onset of menopause compared to controls. This was conducted in UK Biobank and GS:SFHS

254 using unrelated individuals by applying the same criteria as described previously to identify
255 relatedness. In UK Biobank, a linear regression was conducted to compare the age of menopause
256 (Data-Field: 3581) in the CIDI-SF depression cases with controls covarying for the age when attending
257 assessment centre (Data-Field: 21003). Age of attending assessment centre was fitted as a covariate
258 as it was associated with age of menopause (beta = 0.12, standard error = 4.92×10^{-3}, $P = 10^{-137}$).
259 Individuals that reported they had not experienced menopause, were unsure whether menopause
260 had been experienced or had undergone a hysterectomy were excluded. The latest entry for each
261 individual, at either the Initial Assessment (2006-2010), Repeat Assessment (2012-2013) or Imaging
262 visit (2014+), was used to record the age of menopause and age when attending assessment centre.
263 Individuals that had an age at onset of depression (Data-Field: 20433) that was two years prior to or
264 after the age of menopause were classified as controls. In GS:SFHS, a linear regression was also used
265 to compare the age of menopause in MDD cases and controls covarying for age when attending
266 assessment centre. Individuals that had a self-reported age at onset of first episode of MDD [27]
267 obtained during the Structured Clinical Interview that was two years prior to or after the age of
268 menopause were classified as controls. GS:SFHS individuals that reported they had not experienced
269 menopause, had a hysterectomy or whose ovaries had been removed were excluded.

270 **Results**

271 Power analyses of potential subgroups traits

272 To determine whether there was sufficient power (> 0.8) to detect a subgroup and identify the
273 optimum variant selection criterion ($P < 5 \times 10^{-8}$, $P < 10^{-6}$ or $P < 10^{-4}$) for each trait the BUHMBOX
274 power calculation test v0.1 [14] was used. The results of the power analysis for detecting a subgroup
275 for 25 available traits within the two depression definitions are provided in Table 1. Obesity 1 and
276 Obesity 3 were from the same study [35] and were highly correlated (r_g = 0.942, standard error =
277 0.045) and therefore only the trait providing greatest power (Obesity 3) was selected to be tested as

278 a subgroup. The same approach was used for the Squamous Cell Lung Cancer and Lung Cancer traits
279 with only Squamous Cell Lung Cancer selected for analysis.

280 Ten traits had power > 0.8 across both the CIDI-SF depression and broad depression definitions:
281 Schizophrenia [36], Bipolar Disorder [37], Autism Spectrum Disorder [38], Anorexia Nervosa [39],
282 Coronary Artery Disease [40], Crohn's Disease [41], Inflammatory Bowel Disease [42], Obesity 3 [35],
283 Age of Natural Menopause [43], and Squamous Cell Lung Cancer [44]. There was one further trait,
284 Ever Smoked [45], that had power > 0.8 for detection of a subgroup in broad depression.

		CIDI-SF depression		Broad depression	
Subgroup trait	PubMed ID	Optimum variant selection criterion	Power	Optimum variant selection criterion	Power
Neuroticism	24828478	$< 10^{-4}$	0.197	$< 10^{-4}$	0.212
Schizophrenia	25056061	$< 10^{-4}$	**1**	$< 10^{-6}$	**1**
Bipolar Disorder	29906448	$< 10^{-4}$	**1**	$< 10^{-4}$	**1**
Attention Deficit Hyperactivity Disorder	27663945	$< 10^{-4}$	0.319	$< 10^{-4}$	0.416
Autism Spectrum Disorder	28540026	$< 10^{-4}$	**1**	$< 10^{-4}$	**1**
Anorexia Nervosa	28494655	$< 10^{-4}$	**1**	$< 10^{-4}$	**1**
Triglyceride Level	24097068	$< 10^{-4}$	0.265	$< 10^{-4}$	0.364
Coronary Artery Disease	26343387	$< 10^{-4}$	**0.986**	$< 10^{-4}$	**1**
Crohn's Disease	26192919	$< 5 \times 10^{-8}$	**1**	$< 5 \times 10^{-8}$	**1**
Inflammatory Bowel Disease	28067908	$< 5 \times 10^{-8}$	**1**	$< 5 \times 10^{-8}$	**1**
Waist to Hip Ratio	25673412	$< 10^{-4}$	0.114	$< 10^{-4}$	0.117
Body Fat	26833246	$< 10^{-4}$	0.257	$< 10^{-4}$	0.325
Waist Circumference	25673412	$< 10^{-4}$	0.143	$< 10^{-4}$	0.143
Overweight	23563607	$< 10^{-4}$	0.423	$< 10^{-4}$	0.551
Obesity 1	23563607	$< 10^{-4}$	**0.977**	$< 10^{-4}$	**0.996**
Obesity 3	23563607	$< 10^{-4}$	**1**	$< 10^{-4}$	**1**
Body Mass Index	25673413	$< 10^{-4}$	0.135	$< 10^{-4}$	0.14
Age of Menarche	25231870	$< 10^{-4}$	0.461	$< 10^{-4}$	0.537
Age of Natural Menopause	26414677	$< 5 \times 10^{-8}$	**1**	$< 5 \times 10^{-8}$	**1**
Years of Schooling	25201988	$< 10^{-4}$	0.081	$< 10^{-4}$	0.108
College Completion	25201988	$< 10^{-4}$	0.602	$< 10^{-4}$	0.672
Ever Smoked	20418890	$< 10^{-4}$	0.733	$< 10^{-4}$	**0.842**
Age of Smoking Initiation	20418890	$< 10^{-4}$	0.079	$< 10^{-4}$	0.069
Squamous Cell Lung Cancer†	28604730	$< 10^{-5}$	**0.999**	$< 10^{-5}$	**0.999**
Lung Cancer†	28604730	$< 10^{-5}$	**0.813**	$< 10^{-5}$	**0.889**

285 Table 1. Power analysis for detecting a subgroup for 25 traits within either Composite International Diagnostic
286 Interview Short Form (CIDI-SF) depression or broad depression in the UK Biobank discovery cohort. PubMed
287 identifiers (PubMed ID) for the 25 traits are provided. Bold values indicate that power was > 0.8. The optimum

288 variant selection criterion that maximised power for the non-depression traits are provided. †Variants with $P <$
289 10^{-4} were not publicly available for Squamous Cell Lung Cancer or Lung Cancer and so $P < 10^{-5}$ was tested instead.
290

291 Testing for subgroups within depression

292 BUHMBOX v0.38 [14] was used to test ten traits for evidence of a subgroup within CIDI-SF depression
293 and eleven traits within broad depression. The results of the subgroup analyses are provided in Table
294 2. There was evidence of a genetic subgroup relating to age of natural menopause within CIDI-SF
295 depression ($P = 1.69 \times 10^{-3}$) which remained significant after correction for multiple testing. The 47
296 variants used to identify this subgroup are provided in Supplementary Table 3. A genetic subgroup
297 relating to age of menopause was detected within the broad depression phenotype ($P = 9.13 \times 10^{-3}$),
298 although this was not significant after correction for multiple testing.

299 Density plots of the distributions of standardised polygenic risk scores, calculated using 47 variants
300 with $P < 10^{-4}$ for age of natural menopause, in CIDI-SF depression cases and controls with density
301 curves of the estimates for underlying univariate normal distributions are provided in Figure 2. In CIDI-
302 SF depression cases, one normal distribution had a mean polygenic risk score of -0.11 (standard
303 deviation = 0.75) with a second normal distribution with a mean of 0.86 (standard deviation = 0.76).
304 The proportion of individuals in the second normal distribution was 0.11 which is potentially indicative
305 of the proportion of case individuals in the age of menopause subgroup. Cohen's d was greater for the
306 two univariate distributions in CIDI-SF depression cases (1.3) than for the controls (0.5).

Depression definition	Subgroup trait	Number of variants	Depression cases	Depression controls	Subgroup *P*-value
CIDI-SF	Schizophrenia	1,107	1,053	2,393	0.48
	Bipolar Disorder	441	8,027	19,146	0.54
	Autism Spectrum Disorder	56	20,524	49,050	0.94
	Anorexia Nervosa	180	15,233	36,519	0.96
	Coronary Artery Disease	305	12,941	31,186	0.28
	Crohn's Disease	62	23,233	55,940	0.31
	Inflammatory Bowel Disease	146	19,534	46,942	0.22
	Obesity 3	61	22,096	53,312	0.50
	Age of Natural Menopause	47	23,592	56,764	**1.69×10^{-3}**
	Squamous Cell Lung Cancer	52	22,677	54,693	0.48

Broad	Schizophrenia	184	21,451	41,109	0.30
	Bipolar Disorder	440	11,368	21,842	0.59
	Autism Spectrum Disorder	56	29,279	55,882	0.86
	Anorexia Nervosa	180	21,804	41,519	0.95
	Coronary Artery Disease	305	18,514	35,428	0.31
	Crohn's Disease	62	33,137	63,567	0.11
	Inflammatory Bowel Disease	146	27,948	53,235	0.12
	Obesity 3	61	31,690	60,548	0.35
	Age of Natural Menopause	47	33,685	64,533	9.13×10^{-3}
	Squamous Cell Lung Cancer	52	32,535	62,071	0.85
	Ever Smoked	99	27,521	52,296	0.34

Table 2. Evidence of a subgroup from traits tested within either Composite International Diagnostic Interview Short Form (CIDI-SF) depression or broad depression in the UK Biobank discovery cohort. The number of individuals in the UK Biobank discovery cohort classified as depression cases and depression controls is provided. The number of variants assessed is provided based on the optimum variant selection criterion for that trait. Bold values indicate significant evidence of a subgroup after Bonferroni correction for multiple testing.

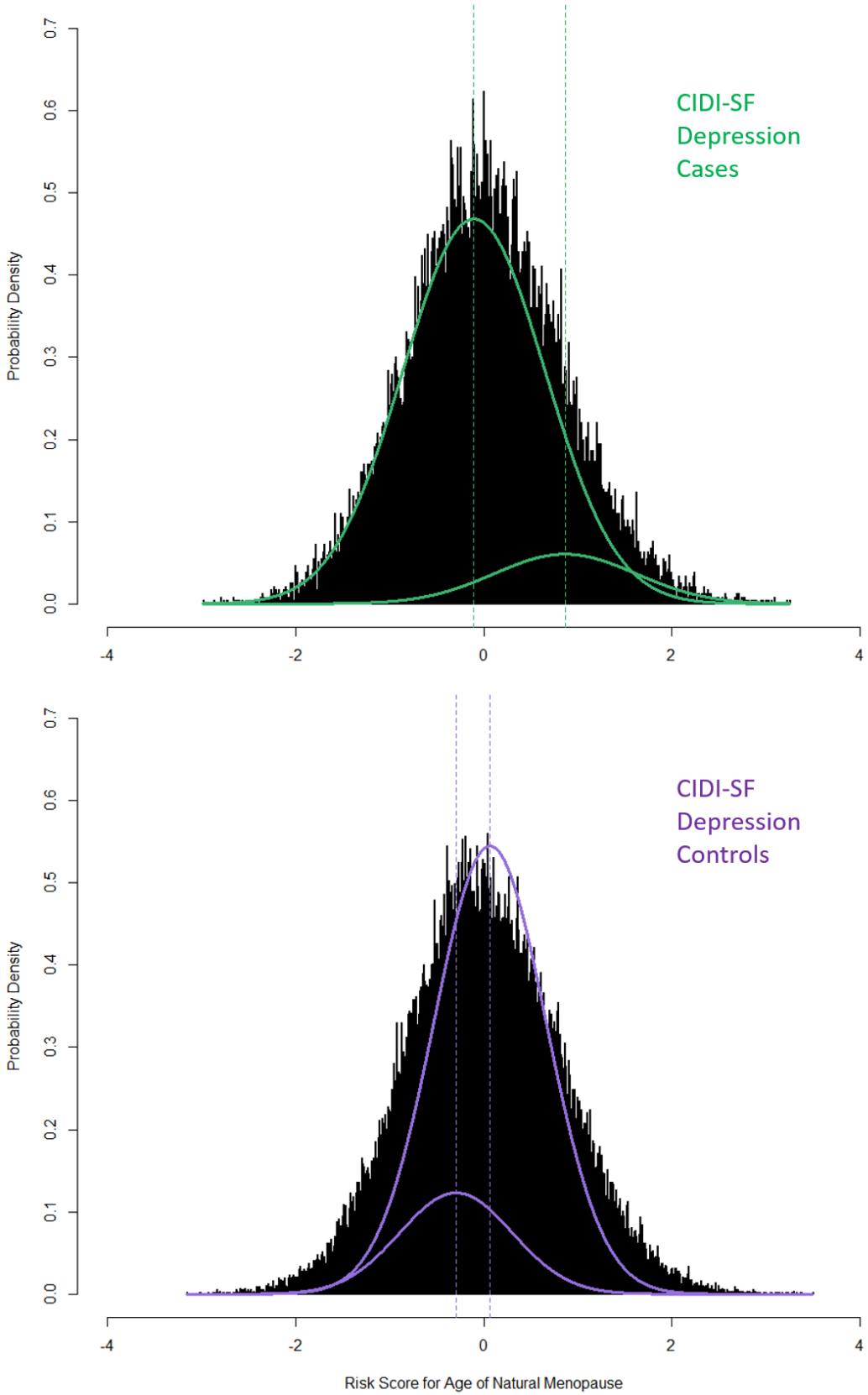

Figure 2. Density plots of the distributions of polygenic risk scores for age of natural menopause in Composite International Diagnostic Interview Short Form (CIDI-SF) depression cases and controls. Overlaid density curves are used to provide estimates of underlying univariate normal distributions in cases (green) and controls (purple).

317 As a subgroup was observed for age of natural menopause, which is a sex-limited trait, the subgroup
318 analysis was rerun in men and woman separately, using variants with $P < 5 \times 10^{-8}$. This would
319 potentially reveal whether it was the genetic variants for age of menopause alone, regardless of sex,
320 which indicated a depression subgroup. In males (7408 cases, 28,558 controls), there was no evidence
321 ($P = 0.186$) for an age of menopause subgroup in CIDI-SF depression. In females (cases = 16,184,
322 controls = 28,206), there remained evidence ($P = 1.18 \times 10^{-3}$) of an age of menopause subgroup within
323 CIDI-SF depression. Using the mix2normal function to examine age of menopause polygenic risk scores
324 for female depression cases estimated one normal distribution with a mean of -0.18 (standard
325 deviation = 0.73) with a second normal distribution with a mean of 0.35 (standard deviation = 0.86)
326 with the proportion of individuals in the first normal distribution estimated as 0.35.

327 To replicate the age of menopause subgroup within CIDI-SF depression observed in the UK Biobank
328 discovery cohort, we also examined the GS:SFHS, iPSYCH, UK Biobank replication and GERA cohorts.
329 There was no evidence of a subgroup for age of natural menopause in any of the replication cohorts
330 ($P \geq 0.05$), when analysing both sexes and in the female only analysis (Table 3). The power was greater
331 in the female only analyses compared to both sexes and this was likely due to the potential subgroup
332 size being larger when analysing only females.

Cohort	Depression definition	Number of variants	Depression cases	Depression controls	Power	Subgroup *P*-value
GS:SFHS	MDD	49	975	5971	0.44	0.47
	MDD (female only)	49	683	3250	0.87	0.74
iPSYCH	Depression	29	19644	21295	0.60	0.24
	Depression (female only)	29	13299	10362	1	0.33
UK Biobank replication	Broad Depression	44	71282	128303	1	0.17
	Broad Depression (female only)	45	45136	59251	1	0.37
GERA	MDD	42	4912	33902	0.92	0.47
	MDD (female only)	42	3543	19494	1	0.63

333 Table 3. *P*-values for an age of natural menopause subgroup in depression within the Generation Scotland:
334 Scottish Family Health Study (GS:SFHS), The Lundbeck Foundation Initiative for Integrative Psychiatric Research
335 (iPSYCH), UK Biobank replication and the Genetic Epidemiology Research on Adult Health and Aging (GERA)
336 cohorts. The number of variants is based on an optimum variant selection criterion of $P < 5 \times 10^{-8}$ for an
337 association with age of natural menopause. Power is based on the estimated proportion of individuals in the age
338 of natural menopause subgroup observed in the UK Biobank discovery cohort (0.11 or 0.35 in female only
339 analysis).

340

341 Phenotypic examination of depression and age of menopause

342 Having observed evidence for a genetic subgroup for age of natural menopause within CIDI-SF
343 depression, we examined whether age of natural menopause differed between depression cases and
344 controls using a linear model. To examine depression prior to onset of menopause the analysis was
345 restricted to cases that reported depression at least two years prior to onset of menopause with age
346 when attending assessment centre (to assess age of menopause) fitted as a covariate in both UK
347 Biobank and GS:SFHS. In UK Biobank, the age of natural menopause in CIDI-SF depression cases (n =
348 7312, mean = 50.24 years) was significantly later (beta = 0.34, standard error = 0.06, $P = 9.92 \times 10^{-8}$)
349 than in controls (n = 21,829, mean = 50.09 years). In GS:SFHS, the age of natural menopause in MDD
350 cases (n = 63, mean = 55.0 years) was earlier than in MDD controls (n = 533, mean 59.0), but after
351 covarying for age of assessment the estimate was in the opposite direction (i.e. depression cases had
352 a later age of menopause) and was not significant (beta = 0.87, standard error = 0.68, $P = 0.20$).

353

354 **Discussion**

355 Depression is a heterogeneous mental health disorder and is comorbid with many other diseases and
356 illnesses. Over the last few years, valuable progress has been made in understanding the underlying
357 genetic architecture of depression [11, 13, 46]. Furthermore, stratifying depression using genetic data
358 remains a key goal within the psychiatric genetics community [47] and should lead to improved
359 classification of mental health conditions and more efficacious treatment for patients. Machine
360 learning [48, 49] and polygenic risk score [6, 50] approaches offer possible methods for stratification
361 in mental health. In the current study, we used BUHMBOX [14] to identify whether traits that were
362 genetically correlated with depression were correlated due to a subgroup, i.e. the correlation was
363 driven by a subset of depressed individuals who had a greater genetic loading for the trait. Evidence
364 of a subgroup within depression may provide future opportunities for stratifying the disease.

365 We examined 25 traits genetically correlated with depression using individuals that had completed
366 the UK Biobank mental health questionnaire. Two definitions of depression were examined to allow a
367 direct comparison between a stricter and a broader definition of depression. We initially conducted a
368 power analysis to determine those correlated traits which could be reasonably tested as genetic
369 subgroups. There were ten traits adequately powered to be tested as subgroups within CIDI-SF
370 depression and eleven traits tested as subgroups within broad depression. A genetic subgroup for age
371 of natural menopause was found within CIDI-SF depression after correction for multiple testing. A
372 genetic subgroup for age of natural menopause was also found within broad depression, but this did
373 not survive multiple testing correction. No evidence for this subgroup was found in GS:SFHS, iPSYCH,
374 a UK Biobank replication cohort or GERA. The lack of replication could be due to Type 1 error, there
375 could be something distinct about the UK Biobank discovery cohort, the different definitions of
376 depression examined, or a combination of factors.

377 From BUHMBOX, it is not directly possible to determine whether an earlier or later age of menopause
378 led to the observed genetic subgroup. However, the phenotypic analyses conducted suggested people
379 with depression have a later age of menopause and so for the purposes of illustrating possible
380 explanations for this subgroup, a later age of menopause is used. Han et al. [14] suggested that
381 subgroups could arise due to ascertainment bias, misclassification, causal relationships, or molecular
382 subgroups. Ascertainment bias seems unlikely as that would require that a later age of menopause
383 somehow increases the chances of individuals receiving clinical attention and obtaining a diagnosis of
384 depression. Misclassification also seems unlikely as there is no obvious reason why individuals with a
385 later age of menopause would be misdiagnosed as depressed. No evidence for a causal relationship
386 in either direction ($P = 0.169$ for depression being causing for age of menopause and $P = 0.529$ for age
387 of menopause being causal for depression) was found by Howard et al. [13] using Mendelian
388 randomization. Molecular subtypes, where there exists a shared developmental pathway between a
389 later age of menopause and depression, represents a potential explanation for our results and
390 identifying this pathway could form the basis of future research.

391 The relationship between depression and menopause has been well studied, but with inconsistent
392 findings [51, 52]. Studies have reported an increase in depressive symptoms during menopause [53-
393 56], but this may be due to the onset of climacteric symptoms, such as insomnia, heavy sweating, hot
394 flashes, and irritability, rather than menopausal state [57-59]. Whereas, Kaufert et al. [60] reported
395 that there was no effect of onset of menopause on depressive status. A meta-analysis of 14 studies
396 found that an older age at menopause led to a lower risk of depression in later life [61]. Several shared
397 neuroendocrine mechanisms have been proposed between menopause and depression. Failure of the
398 gamma-aminobutyric acid A ($GABA_A$) receptor to adapt to fluctuations in ovarian hormones due to the
399 menopause may impact hypothalamic pituitary adrenal (HPA) axis activity [62], with dysregulation of
400 the HPA axis associated with depression [63]. Further, oestradiol is a reproductive hormone that
401 declines during menopause, but it also has a neuroprotective role and contributes to the maintenance
402 of brain homeostasis [64]. A review by Rubinow et al. [65] reported that there was some evidence that
403 oestradiol had an antidepressant effect in perimenopausal women. The role of oestradiol throughout
404 the life course may have produced the results observed in the current study with observable effects
405 on both depression and age of menopause.

406 The results from the subgroup analysis suggest that there was a shared genetic component underlying
407 both depression and age of menopause. Studies examining genetic factors relating to both menopause
408 and mental health phenotypes have principally been focused on the estrogen receptor alpha (*ESR1*)
409 gene [66], with *ESR1* associated with anxiety [67], premenstrual dysphoric disorder [68], and major
410 depressive disorder [69]. However, variants in or near the *ESR1* gene were not associated with age of
411 menopause [43] and therefore not included in the current analysis. Future research identifying genetic
412 factors underlying shared biological mechanisms between menopause and depression may aid in
413 developing new treatments for related mood disorders.

414 The limitations of the current study include selection bias, whereby particular individuals are more
415 likely to participate in population-based cohorts or complete additional assessments, such as the
416 online mental health questionnaire. Participants of the UK Biobank are healthier and from less

417 deprived areas than the general population[70] and those that completed the mental health
418 questionnaire had a lower genetic predisposition to severe depression than those who did not [71].
419 UK Biobank participants that either had a self-reported or a hospital diagnosis of schizophrenia or
420 bipolar disorder were excluded in the current analysis which may limit the potential for identifying
421 subgroups for these disorders. The replication cohorts each used different diagnostic criteria for
422 depression and also examined slightly different sets of genetic variants, nevertheless the set of
423 variants examined were associated with age of menopause. Over half of the traits that are genetically
424 correlated with depression were not included in the subgroup analysis due to a lack of power (≤ 0.8).
425 As increasing large genome-wide association studies become available, a greater number of variants
426 will meet the required selection criteria, allowing additional traits to be tested for evidence of a
427 subgroup within depression.

428 Depression is both polygenic and heterogeneous and stratification of the disorder may lead to
429 improvements in treatment outcomes. In the current study, we found that depressed individuals in
430 the UK Biobank and GS:SFHS had a later age of menopause. This relationship may have a genetic basis
431 with age of natural menopause found to form a subgroup within UK Biobank CIDI-SF depression cases.
432 Using genetic data to identify individuals in this subgroup may ultimately reveal more efficacious
433 treatments for depression.

434

435 Acknowledgements

436 This research was conducted using the UK Biobank resource, application number 4844. We are
437 grateful to the UK Biobank and all its voluntary participants. The UK Biobank study was conducted
438 under generic approval from the NHS National Research Ethics Service (approval letter dated June 17,
439 2011, Ref 11/NW/0382). All participants gave full informed written consent.

440 Generation Scotland received core support from the Chief Scientist Office of the Scottish Government
441 Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. Genotyping of the

442 GS:SFHS samples was funded by the Wellcome Trust (Wellcome Trust Strategic Award "Stratifying
443 Resilience and Depression Longitudinally" (STRADL) Reference 104036/Z/14/Z) and the Medical
444 Research Council UK and was carried out by the Genetics Core Laboratory at the Wellcome Trust
445 Clinical Research Facility, Edinburgh, Scotland. Ethics approval for the Generation Scotland was given
446 by the NHS Tayside committee on research ethics (reference 15/ES/0040), and all participants
447 provided written informed consent for the use of their data.

448 The iPSYCH team acknowledges funding from the Lundbeck Foundation (grants R102-A9118 and R155-
449 2014-1724), the Stanley Medical Research Institute, the European Research Council (project 294838),
450 the Novo Nordisk Foundation for supporting the Danish National Biobank resource, and Aarhus and
451 Copenhagen Universities and University Hospitals, including support to the iSEQ Center, the
452 GenomeDK HPC facility, and the CIRRAU Center. Ethical approval for iPSYCH was provided by the
453 Danish Scientific Ethics Committee, the Danish Health Data Authority, the Danish data protection
454 agency and the Danish Neonatal Screening Biobank Steering Committee.

455 The research conducted using the GERA resource was approved as part of NIH project #21685 and
456 data is available from dbGaP accession reference phs000674.v3.p3. We are grateful to GERA and the
457 Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and its
458 participants. Data came from a grant, the Resource for Genetic Epidemiology Research in Adult Health
459 and Aging (RC2 AG033067; Schaefer and Risch, PIs) awarded to the RPGEH and the UCSF Institute for
460 Human Genetics. The RPGEH was supported by grants from the Robert Wood Johnson Foundation,
461 the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, Kaiser Permanente
462 Northern California, and the Kaiser Permanente National and Northern California Community Benefit
463 Programs.

464 DMH is supported by a Sir Henry Wellcome Postdoctoral Fellowship (Reference 213674/Z/18/Z) and
465 a 2018 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (Ref:
466 27404). A.M.McI acknowledges support from the Wellcome Trust (Wellcome Trust Strategic Award

467 'STratifying Resilience and Depression Longitudinally' (STRADL) Reference 104036/Z/14/Z and The
468 Sackler Trust. N.R.W. acknowledges NMHRC grants 1078901 and 1087889. C.M.L acknowledges MRC
469 grant MR/N015746/1. S.P.H. acknowledges MRC grant MR/S0151132. This investigation represents
470 independent research part-funded by the National Institute for Health Research (NIHR) Biomedical
471 Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The
472 views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the
473 Department of Health.

474 Competing interests

475 Cathryn Lewis is a member of the Science Advisory Board for Myriad Neuroscience. Andrew McIntosh
476 has received speaker fees from Illumina and Janssen. The authors report no other conflicts of interest.

477

478

479 **References**

480 1. Zimmerman, M., W. Ellison, D. Young, I. Chelminski, and K. Dalrymple, *How many different*
481 *ways do patients meet the diagnostic criteria for major depressive disorder?* Comprehensive
482 Psychiatry, 2015. **56**: p. 29-34.
483 2. Fried, E.I., *Moving forward: how depression heterogeneity hinders progress in treatment and*
484 *research.* Expert Review of Neurotherapeutics, 2017. **17**(5): p. 423-425.
485 3. Walker, J., C.H. Hansen, P. Martin, S. Symeonides, R. Ramessur, G. Murray, and M. Sharpe,
486 *Prevalence, associations, and adequacy of treatment of major depression in patients with*
487 *cancer: a cross-sectional analysis of routinely collected clinical data.* The Lancet Psychiatry,
488 2014. **1**(5): p. 343-350.
489 4. Hare, D.L., S.R. Toukhsati, P. Johansson, and T. Jaarsma, *Depression and cardiovascular*
490 *disease: a clinical review.* European Heart Journal, 2013: p. 1365-1372.
491 5. Avenevoli, S., J. Swendsen, J.-P. He, M. Burstein, and K. Merikangas, *Major depression in the*
492 *national comorbidity survey- adolescent supplement: prevalence, correlates, and treatment.*
493 Journal of the American Academy of Child and Adolescent Psychiatry, 2015. **54**(1): p. 37-
494 44.e2.
495 6. Milaneschi, Y., F. Lamers, W.J. Peyrot, A. Abdellaoui, G. Willemsen, J.J. Hottenga, R. Jansen,
496 et al., *Polygenic dissection of major depression clinical heterogeneity.* Molecular Psychiatry,
497 2016. **21**(4): p. 516-522.
498 7. Hall, L.S., M.J. Adams, A. Arnau-Soler, T.-K. Clarke, D.M. Howard, Y. Zeng, G. Davies, et al.,
499 *Genome-wide meta-analyses of stratified depression in Generation Scotland and UK Biobank.*
500 Translational Psychiatry, 2018. **8**(1): p. 9.

8. Adams, M.J., D.M. Howard, M. Luciano, T.-K. Clarke, G. Davies, W.D. Hill, D. Smith, et al., *Genetic stratification of depression by neuroticism: revisiting a diagnostic tradition.* Psychological Medicine, 2019: p. 1-10.
9. Sarda, A., S. Munuswamy, S. Sarda, and V. Subramanian, *Using passive smartphone sensing for improved risk stratification of patients with depression and diabetes: cross-sectional observational study.* JMIR Mhealth Uhealth, 2019. **7**(1): p. e11041.
10. Sullivan, P.F., M.C. Neale, and K.S. Kendler, *Genetic epidemiology of major depression: review and meta-analysis.* American Journal of Psychiatry, 2000. **157**(10): p. 1552-1562.
11. Wray, N.R., S. Ripke, M. Mattheisen, M. Trzaskowski, E.M. Byrne, A. Abdellaoui, M.J. Adams, et al., *Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression.* Nature Genetics, 2018. **50**(5): p. 668-681.
12. Howard, D.M., M.J. Adams, M. Shirali, T.-K. Clarke, R.E. Marioni, G. Davies, J.R.I. Coleman, et al., *Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways.* Nature Communications, 2018. **9**: p. 1470.
13. Howard, D.M., M.J. Adams, T.-K. Clarke, J.D. Hafferty, J. Gibson, M. Shirali, J.R.I. Coleman, et al., *Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions.* Nature Neuroscience, 2019. **22**(3): p. 343-352.
14. Han, B., J.G. Pouget, K. Slowikowski, E. Stahl, C.H. Lee, D. Diogo, X. Hu, et al., *A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases.* Nature Genetics, 2016. **48**(7): p. 803-810.
15. Bycroft, C., C. Freeman, D. Petkova, G. Band, L.T. Elliott, K. Sharp, A. Motyer, et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.
16. Davis, K.A.S., J.R.I. Coleman, M. Adams, N. Allen, G. Breen, B. Cullen, C.M. Dickens, et al., *Mental health in UK Biobank revised.* medRxiv, 2019: p. 19001214.
17. Manichaikul, A., J.C. Mychaleckyj, S.S. Rich, K. Daly, M. Sale, and W.-M. Chen, *Robust relationship inference in genome-wide association studies.* Bioinformatics, 2010. **26**(22): p. 2867-2873.
18. Kessler, R.C., G. Andrews, D. Mroczek, B. Ustun, and H.-U. Wittchen, *The World Health Organization Composite International Diagnostic Interview short-form (CIDI-SF).* International Journal of Methods in Psychiatric Research, 1998. **7**(4): p. 171-185.
19. Chang, C.C., C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, and J.J. Lee, *Second-generation PLINK: rising to the challenge of larger and richer datasets.* GigaScience, 2015. **4**(1): p. 7.
20. Yee, T.W., *The VGAM package for categorical data analysis.* 2010, 2010. **32**(10): p. 34.
21. Smith, B.H., A. Campbell, P. Linksted, B. Fitzpatrick, C. Jackson, S.M. Kerr, I.J. Deary, et al., *Cohort profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness.* International Journal of Epidemiology, 2013. **42**(3): p. 689-700.
22. Nagy, R., T.S. Boutin, J. Marten, J.E. Huffman, S.M. Kerr, A. Campbell, L. Evenden, et al., *Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants.* Genome Medicine, 2017. **9**(1): p. 23.
23. Yang, J., S.H. Lee, M.E. Goddard, and P.M. Visscher, *GCTA: A tool for genome-wide complex trait analysis.* The American Journal of Human Genetics, 2011. **88**(1): p. 76-82.
24. Amador, C., J. Huffman, H. Trochet, A. Campbell, D. Porteous, J.F. Wilson, N. Hastie, et al., *Recent genomic heritage in Scotland.* BMC Genomics, 2015. **16**(1): p. 1-17.
25. Ripke, S. *GWAS genotypic overlap test without sharing genotypes*. 2017; https://personal.broadinstitute.org/sripke/share_links/checksums_download/].
26. First, M.B., R.L. Spitzer, Gibbon Miriam., and J.B.W. Williams, *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition. (SCID-I/P)* 2002.

27. Fernandez-Pujals, A.M., M.J. Adams, P. Thomson, A.G. McKechanie, D.H.R. Blackwood, B.H. Smith, A.F. Dominiczak, et al., *Epidemiology and heritability of major depressive disorder, stratified by age of onset, sex, and illness course in generation scotland: scottish family health study (GS:SFHS).* PLoS ONE, 2015. **10**(11): p. e0142197.
28. O'Connell, J., K. Sharp, N. Shrine, L. Wain, I. Hall, M. Tobin, J.-F. Zagury, et al., *Haplotype estimation for biobank-scale data sets.* Nature Genetics, 2016. **48**(7): p. 817-820.
29. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLOS Genetics, 2009. **5**(6): p. e1000529.
30. The 1000 Genomes Project Consortium., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
31. Pedersen, C.B., J. Bybjerg-Grauholm, M.G. Pedersen, J. Grove, E. Agerbo, M. Bækvad-Hansen, J.B. Poulsen, et al., *The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders.* Molecular Psychiatry, 2017. **23**: p. 6.
32. Kvale, M.N., S. Hesselson, T.J. Hoffmann, Y. Cao, D. Chan, S. Connell, L.A. Croen, et al., *Genotyping informatics and quality control for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort.* 2015. **200**(4): p. 1051-1060.
33. Delaneau, O., J.-F. Zagury, and J. Marchini, *Improved whole-chromosome phasing for disease and population genetic studies.* Nature Methods, 2013. **10**(1): p. 5-6.
34. Banda, Y., M.N. Kvale, T.J. Hoffmann, S.E. Hesselson, D. Ranatunga, H. Tang, C. Sabatti, et al., *Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort.* 2015. **200**(4): p. 1285-1295.
35. Berndt, S.I., S. Gustafsson, R. Magi, A. Ganna, E. Wheeler, M.F. Feitosa, A.E. Justice, et al., *Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture.* Nature Genetics, 2013. **45**(5): p. 501-512.
36. Schizophrenia Working Group of the Psychiatric Genomics Consortium, *Biological insights from 108 schizophrenia-associated genetic loci.* Nature, 2014. **511**(7510): p. 421-427.
37. Ruderfer, D.M., S. Ripke, A. McQuillin, J. Boocock, E.A. Stahl, J.M.W. Pavlides, N. Mullins, et al., *Genomic dissection of bipolar disorder and schizophrenia, Including 28 subphenotypes.* Cell, 2018. **173**(7): p. 1705-1715.e16.
38. Anney, R.J.L., S. Ripke, V. Anttila, J. Grove, P. Holmans, H. Huang, L. Klei, et al., *Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia.* 2017. **8**(1): p. 21.
39. Duncan, L., Z. Yilmaz, H. Gaspar, R. Walters, J. Goldstein, V. Anttila, B. Bulik-Sullivan, et al., *Significant locus and metabolic genetic correlations revealed in genome-wide association study of Anorexia Nervosa.* The American journal of psychiatry, 2017. **174**(9): p. 850-858.
40. The CARDIoGRAMplusC4D Consortium, M. Nikpay, A. Goel, H.-H. Won, L.M. Hall, C. Willenborg, S. Kanoni, et al., *A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease.* Nature Genetics, 2015. **47**: p. 1121.
41. Liu, J.Z., S. van Sommeren, H. Huang, S.C. Ng, R. Alberts, A. Takahashi, S. Ripke, et al., *Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.* Nature Genetics, 2015. **47**(9): p. 979-986.
42. de Lange, K.M., L. Moutsianas, J.C. Lee, C.A. Lamb, Y. Luo, N.A. Kennedy, L. Jostins, et al., *Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease.* Nature Genetics, 2017. **49**: p. 256.
43. Day, F.R., K.S. Ruth, D.J. Thompson, K.L. Lunetta, N. Pervjakova, D.I. Chasman, L. Stolk, et al., *Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair.* Nature Genetics, 2015. **47**: p. 1294.

44. McKay, J.D., R.J. Hung, Y. Han, X. Zong, R. Carreras-Torres, D.C. Christiani, N.E. Caporaso, et al., *Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes.* Nature Genetics, 2017. **49**: p. 1126.
45. The Tobacco Genetics Consortium., *Genome-wide meta-analyses identify multiple loci associated with smoking behavior.* Nature Genetics, 2010. **42**(5): p. 441-447.
46. Hyde, C.L., M.W. Nagle, C. Tian, X. Chen, S.A. Paciga, J.R. Wendland, J.Y. Tung, et al., *Identification of 15 genetic loci associated with risk of major depression in individuals of European descent.* Nature Genetics, 2016. **48**: p. 1031-1036.
47. McIntosh, A.M., P.F. Sullivan, and C.M. Lewis, *Uncovering the genetic architecture of major depression.* Neuron, 2019. **102**(1): p. 91-103.
48. Chekroud, A.M., R.J. Zotti, Z. Shehzad, R. Gueorguieva, M.K. Johnson, M.H. Trivedi, T.D. Cannon, et al., *Cross-trial prediction of treatment outcome in depression: a machine learning approach.* The Lancet Psychiatry, 2016. **3**(3): p. 243-250.
49. Trakadis, Y.J., S. Sardaar, A. Chen, V. Fulginiti, and A. Krishnan, *Machine learning in schizophrenia genomics, a case-control study using 5,090 exomes.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2019. **180**(2): p. 103-112.
50. Wigmore, E.M., J.D. Hafferty, L.S. Hall, D.M. Howard, T.-K. Clarke, C. Fabbri, C.M. Lewis, et al., *Genome-wide association study of antidepressant treatment resistance in a population-based cohort using health service prescription data and meta-analysis with GENDEP.* The Pharmacogenomics Journal, 2019.
51. Llaneza, P., M.P. García-Portilla, D. Llaneza-Suárez, B. Armott, and F.R. Pérez-López, *Depressive disorders and the menopause transition.* Maturitas, 2012. **71**(2): p. 120-130.
52. Vivian-Taylor, J. and M. Hickey, *Menopause and depression: Is there a link?* Maturitas, 2014. **79**(2): p. 142-146.
53. Bromberger, J.T., S.F. Assmann, N.E. Avis, M. Schocken, H.M. Kravitz, and A. Cordal, *Persistent Mood Symptoms in a Multiethnic Community Cohort of Pre- and Perimenopausal Women.* American Journal of Epidemiology, 2003. **158**(4): p. 347-356.
54. Unsal, A., M. Tozun, and U. Ayranci, *Prevalence of depression among postmenopausal women and related characteristics.* Climacteric, 2011. **14**(2): p. 244-251.
55. Cohen, L.S., C.N. Soares, A.F. Vitonis, M.W. Otto, and B.L. Harlow, *Risk for New Onset of Depression During the Menopausal Transition: The Harvard Study of Moods and Cycles.* JAMA Psychiatry, 2006. **63**(4): p. 385-390.
56. Freeman, E.W., *Associations of depression with the transition to menopause.* 2010. **17**(4): p. 823-827.
57. Bosworth, H.B., L.A. Bastian, M.N. Kuchibhatla, D.C. Steffens, C.M. McBride, C. Sugg Skinner, B.K. Rimer, et al., *Depressive symptoms, menopausal status, and climacteric symptoms in women at midlife.* 2001. **63**(4): p. 603-608.
58. Avis, N.E., R. Stellato, S. Crawford, J. Bromberger, P. Ganz, V. Cain, and M. Kagawa-Singer, *Is there a menopausal syndrome? Menopausal status and symptoms across racial/ethnic groups.* Social Science & Medicine, 2001. **52**(3): p. 345-356.
59. Strauss, J.R., *The reciprocal relationship between menopausal symptoms and depressive symptoms: A 9-year longitudinal study of American women in midlife.* Maturitas, 2011. **70**(3): p. 302-306.
60. Kaufert, P.A., P. Gilbert, and R. Tate, *"Reprint of" The Manitoba Project: a re-examination of the link between menopause and depression.* Maturitas, 2008. **61**(1): p. 54-66.
61. Georgakis, M.K., T.P. Thomopoulos, A.-A. Diamantaras, E.I. Kalogirou, A. Skalkidou, S.S. Daskalopoulou, and E.T. Petridou, *Association of age at menopause and duration of reproductive period with depression after menopause: A systematic review and meta-analysis.* JAMA Psychiatry, 2016. **73**(2): p. 139-149.

62. Gordon, J.L., S.S. Girdler, S.E. Meltzer-Brody, C.S. Stika, R.C. Thurston, C.T. Clark, B.A. Prairie, et al., *Ovarian hormone fluctuation, neurosteroids, and HPA axis dysregulation in perimenopausal depression: a novel heuristic model.* The American journal of psychiatry, 2015. **172**(3): p. 227-236.
63. Du, X. and T.Y. Pang, *Is dysregulation of the HPA-axis a core pathophysiology mediating co-morbid depression in neurodegenerative diseases?* 2015. **6**(32).
64. Arevalo, M.-A., I. Azcoitia, and L.M. Garcia-Segura, *The neuroprotective actions of oestradiol and oestrogen receptors.* Nature Reviews Neuroscience, 2014. **16**: p. 17.
65. Rubinow, D.R., S.L. Johnson, P.J. Schmidt, S. Girdler, and B. Gaynes, *Efficacy of estradiol in perimenopausal depression: so much promise and so few answers.* Depression and anxiety, 2015. **32**(8): p. 539-549.
66. Sundermann, E.E., P.M. Maki, and J.R. Bishop, *A review of estrogen receptor alpha gene (ESR1) polymorphisms, mood, and cognition.* Menopause (New York, N.Y.), 2010. **17**(4): p. 874-886.
67. Tiemeier, H., S.C.E. Schuit, T. den Heijer, J.B.J. van Meurs, H.R. van Tuijl, A. Hofman, M.M.B. Breteler, et al., *Estrogen receptor α gene polymorphisms and anxiety disorder in an elderly population.* Molecular Psychiatry, 2005. **10**(9): p. 806-807.
68. Huo, L., R.E. Straub, C. Roca, P.J. Schmidt, K. Shi, R. Vakkalanka, D.R. Weinberger, et al., *Risk for premenstrual dysphoric disorder Is associated with genetic variation in ESR1, the estrogen receptor alpha gene.* Biological Psychiatry, 2007. **62**(8): p. 925-933.
69. Ryan, J., J. Scali, I. Carrière, K. Peres, O. Rouaud, P.-Y. Scarabin, K. Ritchie, et al., *Estrogen receptor alpha gene variants and major depressive episodes.* Journal of Affective Disorders, 2012. **136**(3): p. 1222-1226.
70. Fry, A., T.J. Littlejohns, C. Sudlow, N. Doherty, L. Adamska, T. Sprosen, R. Collins, et al., *Comparison of sociodemographic and health-related characteristics of UK Biobank articipants With those of the general population.* American Journal of Epidemiology, 2017. **186**(9): p. 1026-1034.
71. Adams, M.J., W.D. Hill, D.M. Howard, H.S. Dashti, K.A.S. Davis, A. Campbell, T.-K. Clarke, et al., *Factors associated with sharing e-mail information and mental health survey participation in large population cohorts.* International Journal of Epidemiology, 2019.