The stratification of major depressive disorder into genetic subgroups

David M. Howard*[1], Toni-Kim Clarke [1], Mark J. Adams [1], Jonathan D. Hafferty[1], Eleanor M. Wigmore[1], Yanni Zeng[1, 2], Lynsey S. Hall[1], Jude Gibson[1], Thibaud S. Boutin[2], Caroline Hayward [2, 8], Pippa A. Thomson[3, 4] David J. Porteous[3], Blair H. Smith[5, 8], Alison D. Murray[6, 8], Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium[†], Chris S. Haley[2], Ian J. Deary[4, 7, 8], Heather C. Whalley[1] and Andrew M. McIntosh[1, 7, 8]

Affiliations:
[1]Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK
[2]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[3]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[4]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK
[5]Division of Population Health Sciences, University of Dundee, Dundee, UK
[6]Aberdeen Biomedical Imaging Centre, University of Aberdeen, Aberdeen, UK
[7]Department of Psychology, The University of Edinburgh, Edinburgh, UK
[8]Generation Scotland, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
†Consortium members and their affiliations are listed in supplementary information

*Corresponding author: David M. Howard
Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK
+44 131 537 6268
(e-mail: D.Howard@ed.ac.uk)

Major depressive disorder (MDD) is a heritable condition (h^2 = 37%)[1] and a leading cause of disability worldwide[2]. MDD is clinically heterogeneous and comorbid with a variety of conditions and it has been hypothesised that this causal heterogeneity may have confounded previous attempts to elucidate its genetic architecture[3-5]. We applied a relatively new technique, Buhmbox[6], to identify the presence of heterogeneous sub-groups within MDD using summary data from genome-wide association studies. We analysed two independent cohorts (n_{total} = 31,981) and identified significant evidence ($P_{corrected} < 0.05$) for 10 sub-groups across both cohorts, including subgroups with a liability for migraine, alcohol consumption and eczema. The most notable subgroups ($P_{corrected} \leq 2.57 \times 10^{-8}$ in both cohorts) were for blood levels of cholesterol and triglycerides, and blood pressure, indicating subgroups within MDD cases of individuals with a genetic predisposition for anomalous levels of these metabolic traits. Our findings provide strong evidence for novel causal heterogeneity of MDD and identify avenues for both stratification and treatment.

MDD is a complex and clinically heterogeneous condition that is characterised by symptoms including low mood and/or anhedonia persisting for at least two weeks. Many unique combinations of symptoms may lead to the same diagnosis and it has been suggested that this symptomatic heterogeneity may be due to, as yet unproven, causal heterogeneity[7]. In support of the causal heterogeneity hypothesis, MDD is frequently observed to be comorbid with many diseases including cancer[8], cardiovascular disease,[9] and other psychiatric illnesses[10,11].

We sought to test the presence of causal heterogeneity in MDD according to a number of disease and quantitative traits using a newly available tool, Buhmbox[6]. Buhmbox examines the weighted pairwise correlations of the risk allele dosages for these diseases and traits within MDD cases and controls, based on effect size and frequency, and assigns a *P*-value based on the likelihood of the observed correlations between the cases and controls. We used two cohort studies, Generation Scotland: Scottish Family Health Study (GS:SFHS)[12] and UK Biobank[13], both of which have whole-genome genotyping data and information relating to MDD status. Study demographics for each cohort are provided in **Table 1**. Within each cohort, we examined 34 traits with a reported comorbidity with

MDD and tested whether evidence of subgroups for these traits could be detected within our MDD cases. Further information regarding the 34 traits and their sources is provided in **Supplementary Table 1.** For the traits anorexia nervosa, neuroticism and MDD, summary statistics from different publications were assessed and are numbered accordingly, i.e. MDD 1, MDD 2 and MDD 3. In the case of MDD, this allowed us to examine whether different sets of associated loci, drawn from different populations and diagnostic criteria for MDD, would form a heterogeneous subgroup within our GS:SFHS and/or UK Biobank MDD cases.

Table 1. Study demographics of Generation Scotland: Scottish Family Health Study (GS:SFHS) and UK Biobank

	GS:SFHS	UK Biobank
N	6,946	25,035
Age range	19 - 93	40 - 79
Mean age in years (S.D)	51.5 (13.2)	57.8 (8.0)
Males / Females	3,013 / 3,933	12,528 / 12,507
MDD Cases / Controls	975 / 5,971	8,508 / 16,527

To account for multiple testing, *P*-values were adjusted using a false discovery rate adjustment and all reported values have been adjusted[14]. Ten traits showed significant evidence ($P < 0.05$) of MDD subgroup heterogeneity across both cohorts: total HDL cholesterol levels, total LDL cholesterol levels, serum triglycerides levels, diastolic blood pressure, systolic blood pressure, pulse pressure, alcohol consumption, migraine, eczema and MDD 3. Four traits: Alzheimer's disease, neuroticism, schizophrenia and being a 'morning person' were identified as stratifying traits within one, but not both, cohorts. The subgroup heterogeneity *P*-values obtained within each cohort and for each trait are shown in **Table 2** and **Figure 1**.

Table 2. The false discovery rate adjusted *P*-values for subgroup heterogeneity of the shown disease or quantitative trait within MDD for Generation Scotland: Scottish Family Health Study (GS:SFHS) and UK Biobank. Bold values indicate statistical significance ($P < 0.05$).

	GS:SFHS	UK Biobank
Height	0.657	0.141
Overweight	0.693	0.876
Type 2 Diabetes	0.587	0.854
Alcohol Consumption	**2.29×10^{-4}**	**9.37×10^{-16}**
Cognition	0.387	0.361
Alzheimer's Disease	**0.006**	0.361

Parkinson's Disease	0.851	0.854
Sleep Duration	0.806	0.876
Morning Person	0.300	**0.010**
Anorexia Nervosa 1	0.851	0.781
Anorexia Nervosa 2	0.545	0.588
Migraine	**3.08×10^{-5}**	**5.30×10^{-10}**
Crohn's Disease	0.443	0.606
Inflammatory Bowel Disease	0.651	0.280
Asthma	0.587	0.264
Eczema	**6.56×10^{-3}**	**2.49×10^{-5}**
Rheumatoid Arthritis	0.918	0.174
HDL Cholesterol	**4.09×10^{-14}**	**1.14×10^{-22}**
LDL Cholesterol	**4.91×10^{-16}**	**1.04×10^{-55}**
Triglycerides	**2.57×10^{-8}**	**1.94×10^{-59}**
Diastolic Blood Pressure	**1.48×10^{-26}**	**3.84×10^{-162}**
Systolic Blood Pressure	**1.84×10^{-15}**	**5.07×10^{-27}**
Pulse Blood Pressure	**3.36×10^{-11}**	**1.90×10^{-53}**
Coronary Artery Disease	0.443	0.086
Anxiety Disorder	0.693	0.602
Autism Spectrum Disorder	0.443	0.648
Extraversion	0.784	0.606
Neuroticism 1	0.918	0.102
Neuroticism 2	0.081	**3.20×10^{-15}**
Bipolar Disorder	0.453	0.189
Schizophrenia	0.117	**1.13×10^{-12}**
MDD 1	0.851	0.588
MDD 2	0.851	0.073
MDD 3	**0.023**	**0.004**

88

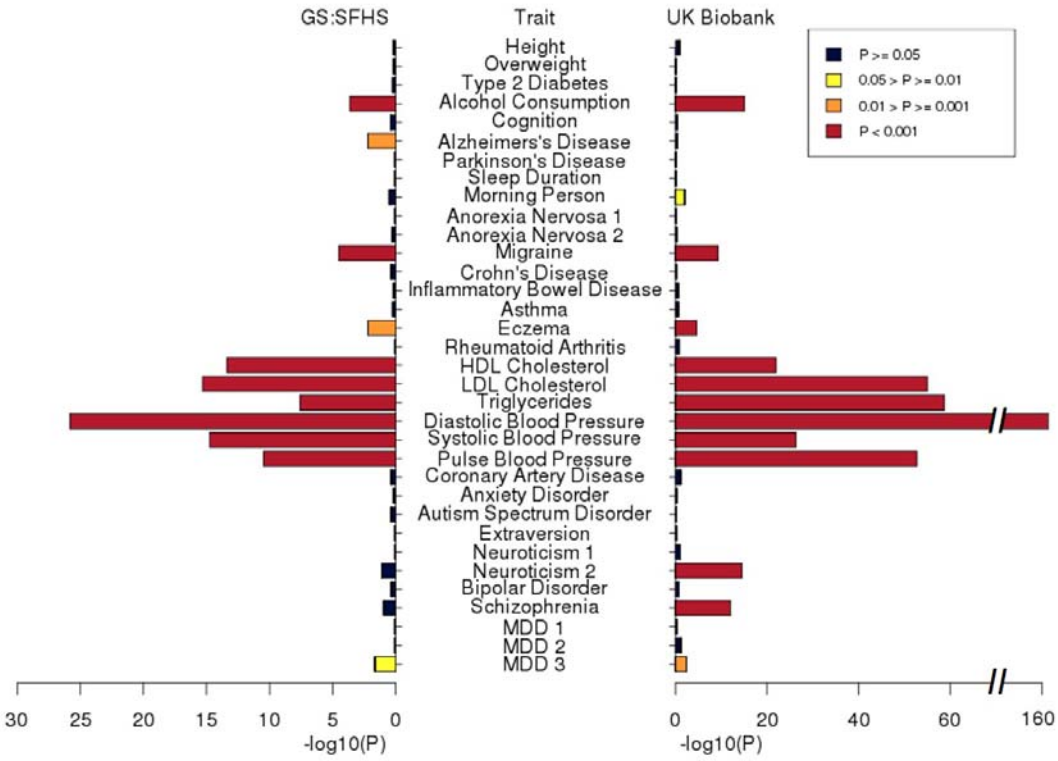

Figure 1. The false discovery rate adjusted *P*-values for evidence of subgroup heterogeneity of the shown disease or quantitative trait within MDD for both Generation Scotland: Scottish Family Health Study (GS:SFHS; n = 6,946) and UK Biobank (n = 25,035).

The most striking results were those relating to metabolic traits ($P \leq 2.57 \times 10^{-8}$ across both cohorts). Although these traits are unlikely to be independent of one another, the statistical significance obtained across both studies suggest robust subgroup heterogeneity. Buhmbox does not identify the individuals within the subgroup and future work to address this would aid in determining the degree of overlap between the observed *P*-values. Milaneschi, et al. [15] also reported evidence of a genetic correlation using profile risk scores between triglycerides and a severe atypical MDD subgroup and it would be beneficial to examine the blood pressure and cholesterol traits in other populations, social, economic and health care settings. Although elevated blood pressure and cholesterol levels are positively correlated with coronary artery disease (CAD), no evidence ($P \geq 0.05$) for a subgroup for CAD was found. This apparent anomaly merits further study, but may reflect stratification independent of CAD.

105 There is substantial comorbidity between alcohol abuse and MDD[16] with studies demonstrating a
106 bidirectional relationship between alcohol and depression[17,18]. Both cohorts used in our study provided
107 evidence ($P < 0.05$) of a novel alcohol consumption subgroup within MDD. Ellingson, et al. [19] have
108 suggested that negative emotionality and behavioral control may mediate the genetic overlap between
109 alcohol consumption and MDD. Further work could examine whether the subgroups within our MDD
110 cases could also possess loci that influence those mediatory factors.

111 Similarly to alcohol consumption, migraine has also been shown to have a bidirectional relationship
112 with depression[20]. We found evidence ($P < 0.05$) of subgroup heterogeneity within MDD that had a
113 genetic predisposition for migraine in both of the cohorts that we studied. Ours is the first study to
114 report the existence of a heterogeneous subgroup within MDD cases of individuals with a migraine-
115 like genetic profile.

116 Subgroup heterogeneity for MDD was observed ($P < 0.05$) in both GS:SFHS and UK Biobank for the
117 MDD 3 trait obtained from Hyde, et al. [21]. MDD 3 was based on a self-reported diagnosis of MDD
118 within a large and predominately European population. The diagnosis of MDD within UK Biobank
119 was also self-reported and the existence of an MDD 3 subgroup suggests that there is a common
120 shared genetic basis to this phenotype, but that it was not shared across all cases. MDD 1 was
121 extracted from a study examining recurrent depression in Han Chinese women within a hospital
122 setting and although we didn't find evidence of a subgroup ($P \geq 0.05$), this was not completely
123 unexpected within our population and UK-based cohorts. No evidence ($P \geq 0.05$) was found for a
124 MDD 2 subgroup, however a polygenic risk score approach has provided evidence of pleiotropy
125 between MDD 2 cases and GS:SFHS cases ($P < 1.37 \times 10^{-10}$) and MDD 2 cases and UK Biobank
126 cases ($P < 1.92 \times 10^{-8}$), Hall, et al., (manuscript in preparation).

127 The body's inflammatory response has been highlighted as a potential contributor to depression[22].
128 Crohn's disease, inflammatory bowel disease and asthma were examined, but neither cohort provided
129 evidence ($P \geq 0.05$) for subgroup heterogeneity. Asthma is frequently comorbid with eczema and both
130 cohorts examined in our study provided evidence ($P < 0.05$) for a subgroup of individuals with a
131 genetic predisposition for eczema within our MDD cases. Associations between eczema and

depression are well reported in the literature and potentially mediated by health anxiety[23]. Chronic inflammation seen in some cases of eczema, and a growing appreciation for the impotence of inflammation in depression, provides another possible explanation for subgroup heterogeneity.

A degree of cognitive impairment has been reported in individuals that are currently experiencing a depressive episode[24,25]. However, we found no evidence ($P \geq 0.05$) for MDD subgroup heterogeneity for general fluid cognitive ability. Alzheimer's disease is also associated with a decline in cognitive ability and previous studies have demonstrated depression to be a risk factor for the disease[26,27]. Within GS:SFHS, there was evidence ($P = 0.006$) for a subgroup of MDD cases which harboured the loci associated with Alzheimer's disease, however this was not replicated in UK Biobank ($P \geq 0.05$). Parkinson's disease is another condition that is associated with neuropathology that is more likely to occur later in life. We found no evidence ($P \geq 0.05$) in either cohort for a subgroup of Parkinson's disease within MDD cases.

Subgroup heterogeneity was observed ($P = 1.13 \times 10^{-12}$) in UK Biobank for schizophrenia which substantiates the work of Milaneschi, et al. [7] who demonstrated correlated genomic profile risk scores between schizophrenia and a severe typical MDD subtype. However, Han, et al. [6] and our GS:SFHS cohort provided no evidence ($P > 0.05$) of a schizophrenia subgroup within MDD cases, which suggests that evidence of subgroup heterogeneity for schizophrenia is population and/or diagnosis dependent.

We also examined a number of developmental and personality traits due to the impact that depression can have on social interaction and feelings of self-worth. Evidence of subgroup heterogeneity ($P = 3.20 \times 10^{-15}$) was only found within UK Biobank for neuroticism 2 drawn from the Smith, et al. [28] study. The neuroticism 2 trait had a much greater number of associated loci compared to neuroticism 1 and therefore neuroticism 1 may have been underpowered to detect an effect, but this is also dependent on the effect sizes of the associated loci, the number of MDD cases and the size of any subgroup. There are similarities in the way that individuals respond to stressful events between neuroticism and MDD and it may be the heritable component that underpins this response that is driving the observed subgrouping within UK Biobank.

The diagnosis of MDD within GS:SFHS was based on the DSM-IV criteria[29], which includes questions related to sleep and eating patterns. Therefore, sleep duration, being a 'morning person' and anorexia nervosa were included in our study. However, it was only within UK Biobank that a significant *P*-value for being a 'morning person' ($P = 0.010$) was observed. Anorexia nervosa 1 and 2 and sleep duration had low numbers of associated loci available (**Supplementary Table 1**) and were potentially underpowered to detect an effect.

Rheumatoid Arthritis (RA) is an autoimmune disease which, like many other chronic diseases, has been shown in multiple studies to be comorbid with depression[30,31], with a potential subgroup of depressed individuals within RA sufferers[32]. However, no evidence ($P \geq 0.05$) of subgroup heterogeneity was found for RA within either GS:SFHS or UK Biobank MDD cases. Multiple studies have suggested that morphological traits[33,34] and type 2 diabetes[35] may identify subgroups of individuals with MDD, but we found no evidence ($P \geq 0.05$) for subgroup heterogeneity according to these traits in the current study.

The two cohorts used in this study reflect a subsample of the UK population, with additional steps taken to ensure that were no overlapping individuals. An MDD diagnosis was made using a structured clinical interview within GS:SFHS, whereas UK Biobank cases were defined by a number of self-reported measures. A broad range of traits were assessed and the selection of the summary statistics used was based on the number of individuals analysed, the availability of summary statistics and publication date. Buhmbox measures the correlations within cases, which is independent (or orthogonal) information from the effect size (personal communication with Buhm Han). Therefore, UK Biobank was able to be used to obtain the summary statistics for neuroticism 2, alcohol consumption and the blood pressure traits, and then also used to assess the existence of MDD subgroup heterogeneity.

Multiple studies have suggested the presence of aetiology subgroups within MDD and Buhmbox provides a quantifiable measure of their existence. Our study has provided replicable evidence of novel subgroup heterogeneity within MDD for a range of disease and quantitative traits, including blood pressure, cholesterol and triglyceride levels, migraine, eczema and alcohol consumption. This

research underlines the potential of using genomic data for developing stratified approaches to the diagnosis and treatment of depression.

COMPETING FINANCIAL INTERESTS

The authors declare that no competing financial interests exist.

ACKNOWLEDGEMENTS

Please refer to the supplementary information for full acknowledgments.

AUTHOR CONTRIBUTIONS

AMMcI, DJP, BHS, ADM, IJD and CH were involved in the acquisition of the GS:SFHS cohort. Quality control of the GS:SFHS data was conducted by LSH, JDH, MJA, CH and DHM. Imputation of the GS:SFHS data was conducted by TB and CH. Quality control of the UK Biobank data was conducted by MJA and DHM. AMMcI and DMH conceived the initial design of the study with HCW, JDH, YZ, T-KC, MJA, EMW, JG, PAT, CSH, IJD and DJP involved in the ongoing development of the project. DMH conducted the analysis and wrote the paper and all authors have read and approved its submission.

209

210

211 Methods

212 Generation Scotland: Scottish Family Health Study (GS:SFHS)

213 The family and population-based Generation Scotland: Scottish Family Health Study (GS:SFHS)
214 cohort[12] consisted of 23,960 individuals, of whom 20,195 were genotyped with the Illumina
215 OmniExpress BeadChip (706,786 SNPs). The genotypic data was uploaded to the Michigan
216 Imputation Server[36] and phased using SHAPE IT v2.r837[37] and imputed using the Haplotype
217 Reference Consortium reference panel (HRC.r1-1)[38]. The imputation of GS:SFHS has been published
218 previously[39]. We applied an imputation accuracy threshold (infoscore) of $\geq$ 0.8 and this provided us
219 with a total of 8,633,288 genome-wide variants calls for 20,032 individuals.

220 A diagnosis of MDD was made using two initial screening questions and the Structured Clinical
221 Interview for the Diagnostic and Statistical Manual of Mental Disorders (SCID)[29]. The diagnosis of
222 MDD within GS:SFHS has been described previously[40] and in our study, MDD was defined by at
223 least one instance of a major depressive episode. Further to this, we used record linkage to the
224 Scottish Morbidity Record[41] to examine the psychiatric history of both case and control individuals.
225 We identified 1,072 control individuals who had attended at least one psychiatry outpatient clinic and
226 we excluded these individuals from our study. Using the psychiatric inpatient records, we identified
227 47 MDD cases who were also diagnosed with bipolar disorder or schizophrenia and these individuals
228 were also excluded from our study. These participants provided us with prior consent for their
229 anonymised data to be linked to clinical data. As GS:SFHS was a family-based cohort, we created an
230 unrelated subsample using GCTA v.122[42] ensuring that no two individuals shared a genomic
231 relatedness of $\geq$ 0.025. A further 186 individuals who were identified as population outliers through
232 principal component analyses of their genotypic information[43]. This left a total of 975 MDD cases and
233 5,971 controls (14.0% prevalence) in the GS:SFHS cohort.

234 UK Biobank

The population-based UK Biobank[13] (provided as part of project #4844) consisted of 152,249 individuals with genomic data for 72,355,667 imputed variants[44]. This was the standard data release available to all approved researchers of UK Biobank. Detailed information regarding the imputation procedure[45] and initial quality control[46] are provided elsewhere. In summary, phasing was achieved using a modified version of SHAPE IT 2[47] with a combined reference panel of 1,000 genomes phase 3 and the UK10K haplotype reference[48] panels and the IMPUTE2 package[49] used for imputation. We applied an infoscore threshold of $\geq$ 0.8 which left a total of 24,467,210 variants. We removed individuals listed as non-white British and those individuals that had also participated in GS:SFHS identified using a checksum approach[50] using genotype data.

Of the remaining participants, 25,035 had completed a touchscreen assessment of depressive symptoms and previous treatment. We used the diagnostic definitions of Smith, et al. [51] and defined case status as either 'probable single lifetime episode of major depression' or 'probable recurrent major depression (moderate and severe)' and with control status defined as 'no mood disorder'. This provided us with a total of 8,508 cases and 16,527 controls (34.0% prevalence) within UK Biobank, which is greater than that observed within GS:SFHS.

Statistical Approach

Buhmbox v0.33[6] was used to conduct the statistical analysis and this package requires raw genetic and phenotypic data (disease A) and also summary statistics relating to the additional disease and quantitative traits for testing (disease B). The disease B associated loci were drawn from either published material or from personal communications and are detailed in **Supplementary Table 1**. The pruning of disease B associated loci was conducted using Plink 1.90[52] and the --indep-pairwise command. A 50 variant window with a 5 variant sliding window was applied to the summary statistics and pruned any variants with an $r^2 > 0.1$.

For GS:SFHS the first 20 principal components were derived from the genotypic data using GCTA v1.22[42] and these were fitted within Buhmbox to account for population stratification. For UK Biobank the first 15 genetic principal components[53] were fitted. Buhmbox examines whether there is a

sharing of risk alleles between the disease B associated loci and the disease A cases (in our case MDD). Buhmbox uses the positive correlations between risk allele dosages in disease A cases to determine whether any sharing of risk alleles is driven by all individuals (pleitropy) or by a subset of individuals (heterogeneity). The likelihood of observing such positive correlations are used to determine the reported *P*-values. The Buhmbox software and manual is freely downloadable from http://software.broadinstitute.org/mpg/buhmbox/. The data that support the findings of this study are available on reasonable request from the corresponding author, DMH. The data are not publicly available due to participant confidentiality and the terms of the existing mutual transfer agreements with the respective data repositories.

1. Sullivan, P.F., Neale, M.C. & Kendler, K.S. Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry* **157**, 1552-1562 (2000).
2. World Health Organization. Mental health action plan 2013 - 2020. (2013).
3. Flint, J. & Kendler, Kenneth S. The genetics of major depression. *Neuron* **81**, 484-503 (2014).
4. Levinson, D.F. *et al.* Genetic studies of major depressive disorder: why are there no genome-wide association study findings and what can we do about it? *Biological Psychiatry* **76**, 510-512 (2014).
5. Lubke, G.H. *et al.* Genome-wide analyses of borderline personality features. *Molecular Psychiatry* **19**, 923-929 (2014).
6. Han, B. *et al.* A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. *Nature Genetics* **48**, 803-810 (2016).
7. Milaneschi, Y. *et al.* Polygenic dissection of major depression clinical heterogeneity. *Molecular Psychiatry* **21**, 516-522 (2016).
8. Walker, J. *et al.* Prevalence, associations, and adequacy of treatment of major depression in patients with cancer: a cross-sectional analysis of routinely collected clinical data. *The Lancet Psychiatry* **1**, 343-350 (2014).
9. Hare, D.L., Toukhsati, S.R., Johansson, P. & Jaarsma, T. Depression and cardiovascular disease: a clinical review. *European Heart Journal*, 1365-1372 (2013).
10. Häfner, H., Löffler, W., Maurer, K., Hambrecht, M. & an der Heiden, W. Depression, negative symptoms, social stagnation and social decline in the early course of schizophrenia. *Acta Psychiatrica Scandinavica* **100**, 105-118 (1999).
11. Avenevoli, S., Swendsen, J., He, J.-P., Burstein, M. & Merikangas, K. Major depression in the national comorbidity survey- adolescent supplement: prevalence, correlates, and treatment. *Journal of the American Academy of Child and Adolescent Psychiatry* **54**, 37-44.e2 (2015).
12. Smith, B.H. *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* **42**, 689-700 (2013).

13. Allen, N.E., Sudlow, C., Peakman, T. & Collins, R. UK biobank data: come and get it. *Science Translational Medicine* **6**, 224ed4 (2014).
14. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
15. Milaneschi, Y. *et al.* Polygenic dissection of major depression clinical heterogeneity. *Mol Psychiatry* **21**, 516-522 (2016).
16. Grant, B.F., Stinson, F.S., Dawson, D.A. & et al. Prevalence and co-occurrence of substance use disorders and independentmood and anxiety disorders: Results from the national epidemiologic survey on alcohol and relatedconditions. *Archives of General Psychiatry* **61**, 807-816 (2004).
17. Brière, F.N., Rohde, P., Seeley, J.R., Klein, D. & Lewinsohn, P.M. Comorbidity between major depression and alcohol use disorder from adolescence to adulthood. *Comprehensive Psychiatry* **55**, 526-533 (2014).
18. Flensborg-Madsen, T. *et al.* Comorbidity and temporal ordering of alcohol use disorders and other psychiatric disorders: results from a Danish register-based study. *Comprehensive Psychiatry* **50**, 307-314 (2009).
19. Ellingson, J.M., Richmond-Rakerd, L.S., Statham, D.J., Martin, N.G. & Slutske, W.S. Most of the genetic covariation between major depressive and alcohol use disorders is explained by trait measures of negative emotionality and behavioral control. *Psychological Medicine* **46**, 2919-2930 (2016).
20. Breslau, N., Lipton, R.B., Stewart, W.F., Schultz, L.R. & Welch, K.M.A. Comorbidity of migraine and depression: Investigating potential etiology and prognosis. *Neurology* **60**, 1308-1312 (2003).
21. Hyde, C.L. *et al.* Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nature Genetics* **48**, 1031-1036 (2016).
22. Miller, A.H. & Raison, C.L. The role of inflammation in depression: from evolutionary imperative to modern treatment target. *Nat Rev Immunol* **16**, 22-34 (2016).
23. Klokk, M., Gotestam, K.G. & Mykletun, A. Factors accounting for the association between anxiety and depression, and eczema: the Hordaland health study (HUSK). *BMC Dermatology* **10**, 3 (2010).
24. Millan, M.J. *et al.* Cognitive dysfunction in psychiatric disorders: characteristics, causes and the quest for improved therapy. *Nat Rev Drug Discov* **11**, 141-168 (2012).
25. Dere, E., Pause, B.M. & Pietrowsky, R. Emotion and episodic memory in neuropsychiatric disorders. *Behavioural Brain Research* **215**, 162-171 (2010).
26. Diniz, B.S., Butters, M.A., Albert, S.M., Dew, M.A. & Reynolds, C.F. Late-life depression and risk of vascular dementia and Alzheimer's disease: systematic review and meta-analysis of community-based cohort studies. *The British Journal of Psychiatry* **202**, 329-335 (2013).
27. Norton, S., Matthews, F.E., Barnes, D.E., Yaffe, K. & Brayne, C. Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. *The Lancet Neurology* **13**, 788-794 (2014).
28. Smith, D.J. *et al.* Genome-wide analysis of over 106 000 individuals identifies 9 neuroticism-associated loci. *Molecular Psychiatry* **21**, 749-757 (2016).
29. First, M.B., Spitzer, R.L., Gibbon Miriam. & Williams, J.B.W. Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition. (SCID-I/P) (2002).
30. Matcham, F., Rayner, L., Steer, S. & Hotopf, M. The prevalence of depression in rheumatoid arthritis: a systematic review and meta-analysis. *Rheumatology* **52**, 2136-2148 (2013).
31. Rathbun, A.M., Harrold, L.R. & Reed, G.W. Temporal associations between the different domains of rheumatoid arthritis disease activity and the onset of patient-reported depressive symptoms. *Clinical Rheumatology* **34**, 653-663 (2015).

32. Tillmann, T., Krishnadas, R., Cavanagh, J. & Petrides, K. Possible rheumatoid arthritis subtypes in terms of rheumatoid factor, depression, diagnostic delay and emotional expression: an exploratory case-control study. *Arthritis Research & Therapy* **15**, R45 (2013).
33. Deaton, A. & Arora, R. Life at the top: The benefits of height. *Economics & Human Biology* **7**, 133-136 (2009).
34. Carey, M. *et al.* Prevalence of comorbid depression and obesity in general practice: a cross-sectional survey. *British Journal of General Practice* **64**, e122-e127 (2014).
35. Rustad, J.K., Musselman, D.L. & Nemeroff, C.B. The relationship of depression and diabetes: Pathophysiological and treatment implications. *Psychoneuroendocrinology* **36**, 1276-1286 (2011).
36. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nature Genetics* **48**, 1284-1287 (2016).
37. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10**, 5-6 (2013).
38. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48**, 1279-1283 (2016).
39. Nagy, R. *et al.* Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Medicine* **9**, 23 (2017).
40. Fernandez-Pujals, A.M. *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* **10**, e0142197 (2015).
41. Information Services Division. SMR Data Manual. (http://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets, 2016).
42. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**, 76-82 (2011).
43. Amador, C. *et al.* Recent genomic heritage in Scotland. *BMC Genomics* **16**, 1-17 (2015).
44. Marchini, J. UK Biobank phasing and imputation documentation. Version 1.2. (http://biobank.ctsu.ox.ac.uk/crystal/docs/impute_ukb_v1.pdf, 2015).
45. UK Biobank. Genotype imputation and genetic association studies using UK Biobank data. http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=157020 (2015).
46. UK Biobank. Genotyping and quality control for UK Biobank. http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155580 (2015).
47. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nature Genetics* **48**, 817-820 (2016).
48. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature Communications* **6**, 8111 (2015).
49. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3: Genes|Genomes|Genetics* **1**, 457-470 (2011).
50. Ripke, S. GWAS genotypic overlap test without sharing genotypes. https://personal.broadinstitute.org/sripke/share_links/checksums_download/readme_0415b.txt (2014).
51. Smith, D.J. *et al.* Prevalence and characteristics of probable major depression and bipolar disorder within UK Biobank: cross-sectional study of 172,751 participants. *PLoS ONE* **8**, e75362 (2013).
52. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
53. UK Biobank. Genetic principal components. http://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=22009 (2015).