**Genetic and clinical characteristics of treatment-resistant depression using primary care records in two UK cohorts**

**Supplementary Methods**

**1. Primary care data in UK Biobank (UKB)**

There is currently no national system for collecting or sharing primary care data in the UK. UKB has been liaising with various data suppliers and other intermediaries (including the main primary care computer system suppliers in England) to obtain primary care data for UKB participants, all of whom have provided written consent for linkage to their health-related records. To date, coded data have been obtained for approximately 45% of the UK Biobank cohort (~230,000 participants) and are now available as part of this interim release. UKB is currently in the process of securing access to data for the remaining cohort, mainly for participants registered with EMIS (a computer system supplier to the NHS) practices across England (UK Biobank, 2019).

The dataset contains variables that are considered the most important for epidemiological research: coded clinical events (including diagnoses, history, symptoms, lab results, procedures), prescriptions (i.e. medications that are prescribed but not necessarily dispensed) and a range of administrative codes (e.g. referrals to specialist hospital clinics). Non-coded, unstructured data (e.g. free-text entries, referral letters) are not included (UK Biobank, 2019).

The primary care computer system suppliers have adopted different coding classifications as part of their underlying data schema. In addition to these coding variations for clinical events, the different system suppliers use a range of coding classifications for prescriptions, as reported in the following table:

| **Country** | **GP Computer System Supplier** | **Approx no. of UKB participants** | **Clinical coding classification** | **Prescription coding classification** |
|---|---|---|---|---|
| Scotland | EMIS / Vision | 27,000 | Read v2 | - Read v2<br>- British National Formulary (BNF) |
| Wales | EMIS / Vision | 21,000 | Read v2 | Read v2 |
| England | TTP | 165,000 | Clinical Terms Version 3 (CTV3 or Read v3) | BNF |
| | Vision | 18,000 | Read v2 | - Read v2<br>- Dictionary of Medicines and Devices (dm+d) |

Read codes are a coded thesaurus of clinical terms used in primary care since 1985. There are two versions: version 2 (Read v2) and version 3 (CTV3 or Read v3). Both provide a standard vocabulary for clinicians to record patient findings and procedures. Read v2 and CTV3, together with a UK Read code browser, are available via the NHS Digital Technology Reference Data Update Distribution

(TRUD) website4. Read v2 and CTV3 were last updated in April 2016 and April 2018, respectively. Both versions are now deprecated and no further updates will occur. From April 2018, SNOMED CT was introduced into primary care in a phased approach and it is intended by April 2020 that SNOMED CT will be fully incorporated across the wider NHS, including codes related to prescriptions (UK Biobank, 2019).

The BNF is the standard list of medicines, dressings and appliances prescribed in the UK. It is published as a reference guide in both online and paper versions and contains information on, for example, dose, side effects and price for over 70,000 items. Code lists are updated annually and can be downloaded from the NHS Business Services Authority (NHSBSA) (UK Biobank, 2019).

The prescription data from Vision (England) contains dm+d codes (as well as Read v2 codes) to record medicines prescribed to patients. The dm+d dictionary has been developed for use throughout the NHS (primary and secondary care) to identify specific medicines and devices used in the treatment of patients and consists of a dictionary containing unique identifiers and associated text descriptions (UK Biobank, 2019).

In order to facilitate research on these data, clinical code lists have been compiled from TRUD and NHSBSA (Appendix C) [see Resource 592]. TRUD has historically provided information on how to map from Read v2 and CTV3 to other clinical coding systems (UK Biobank, 2019).

Information on participant registrations varies by data supplier, in that Vision (England) provided a single registration record per person while the other suppliers provided multiple records per participant, and a small number of participants with data in the TPP extract do not have a registration record. Therefore, variable numbers of registration records are included in this release, reflecting the providers' extracts. The start date of coverage is not known for all participants, nor is the completeness of coverage of their primary care health records until the extract date (UK Biobank, 2019).

Data were extracted from supplier's computer systems using different approaches, in each case making a single extract date or cut-off point impossible to determine. For more information see (UK Biobank, 2019).

We classified Read v2 and CTV3 clinical codes in diagnostic groups (e.g. depressive disorders, bipolar disorders, anxiety disorders) and we linked Read v2 clinical codes to the corresponding CTV3 clinical codes (Supplementary Tables 1 and 2). Prescription codes were reported not only according to different classifications but within the same classification system different formats were used (e.g. BNF codes sometimes included dots while other times they didn't), therefore we manually inspected samples of the data to find irregularities and extract the records of interest accordingly. For some prescription records the only information provided was medication code and issue date, though in most cases also the drug name was included as reported in the drug label (brand or generic). Therefore, we had to annotate prescription records with medication chemical name and class (e.g. antidepressant, antipsychotic) using as reference information provided by NHS websites (dm+d browser; British National Formulary). These annotated tables are in Supplementary Tables 1 and 2 in order to facilitate the extraction of data by other investigators.

Where clinical event or prescription date preceded or matched participant date of birth, it was in the year of their birth, or it was in the future, it has been altered to some predefined values in UKB

data (01/01/1901, 02/02/1902, 03/03/1903 and 07/07/2037), and these values were set to missing for the analyses of this study.

**2. Other measures of depression in UKB**

Five other measures of depression were considered in UKB for comparison with primary care-defined depression (Figure 1):

1) Lifetime depression defined based on the Composite International Diagnostic Interview Short Form (CIDI-SF) (Kessler et al., 1998) that was part of the Mental Health Questionnaire (MHQ). Criteria for lifetime major depressive episode were in accordance with DSM-V. The full CIDI is a validated measure of depression, demonstrated to have good concordance with direct clinical assessment (Haro et al., 2006).
2) Lifetime depression based on hospital diagnosis (ICD-10 codes F32-F33-F34-F38-F39), considering both main ICD-10 diagnoses (data field 41202) and secondary ICD-10 diagnoses (data field 41204). Individuals having at least one ICD-10 code but no ICD-10 code for a depressive disorder were considered as having no lifetime depression based on this measure.
3) Self-reported depression diagnosed by a professional (data field 20544). This corresponded to the question: "Have you been diagnosed with one or more of the following mental health problems by a professional, even if you don't have it currently?".
4) Help-seeking definition of depression according to the questions "Have you ever seen a general practitioner (GP) for nerves, anxiety, tension or depression?" (data field 2090) and "Have you ever seen a psychiatrist for nerves, anxiety, tension or depression?" (data field 2100). Individuals who answered 'yes' to at least one of these two questions were considered as having a current or lifetime depression.
5) Smith et al. definition of depression (Smith et al., 2013). This was defined as probable lifetime depression based on a combination of measures (items relating to the lifetime experience of minor and major depression, items from the Patient Health Questionnaire (PHQ) and items on help-seeking for mental health). In detail, this phenotype was defined as having satisfied at least one of the following two groups of criteria: a) ever felt depressed/down for a whole week; plus at least two weeks duration of depression; plus ever seen a GP or psychiatrist for "nerves, anxiety, tension or depression"; OR b) ever anhedonic for a whole week; plus at least two weeks duration of depression; plus ever seen a GP or psychiatrist for "nerves, anxiety, tension or depression".

**3. Genotyping, quality control and imputation**

**3.1. UK biobank**

Autosomal genotype data underwent centralised quality control to adjust for possible array effects, batch effects, plate effects, and departures from Hardy-Weinberg equilibrium (HWE) (Bycroft et al., 2018). SNPs were further excluded based on missingness (> 0.02) and on Hardy Weinberg equilibrium ($p < 10\text{-}8$). Individuals were removed for high levels of missingness (> 0.05) or abnormal heterozygosity (as defined during centralised quality control), relatedness of up to third-degree

kinship (KING r < 0.044 (Manichaikul et al., 2010)) or phenotypic and genotypic gender discordance. Population structure within the UK Biobank cohort was assessed using principal component analysis, with European ancestry defined by 4-means clustering on the first two genetic principal components (Warren et al., 2017).

### 3.2. EXCEED cohort

Variants for the polygenic risk score analysis were limited to common variants (minor allele frequency >0.01) that were directly genotyped. Variants were further excluded based on missingness (>0.05) and on Hardy-Weinberg equilibrium $P < 1\times10^{-6}$. Variants were checked for plate effects, i.e. variants that have significantly different minor allele counts on a particular plate compared to all other plates. A $\chi^2$ test was used to compare the minor allele count of a variant on each plate to its minor allele count on all other plates. A *P*-value of $1\times10^{-12}$ was used to indicate a significant plate effect, and variants were excluded or set to missing according to how many plates a variant showed a significant plate effect.

Individuals were excluded for high levels of missingness (>0.05) or abnormal ancestry-adjusted heterozygosity rate (more than 6 standard deviations from the mean), as conducted by UK Biobank (Bycroft et al., 2018), relatedness of up to third-degree kinship (according to IBD analysis in PLINK 1.90 (Chang et al., 2015) (PI_HAT >0.125), the individual with the highest missingness was excluded from each related pairing), or phenotypic and genotypic gender discordance. Population structure was assessed using principal components analysis. The starting cluster centres for *k*-means clustering were defined as the mean principal components for each of the 5 1000 Genomes Phase 3 super populations (EUR, AFR, EAS, SAS, AMR). This was done as unsupervised *k*-means clustering (cluster centres randomly selected) was unsuccessful for the EXCEED samples. European ancestry was, thus, defined by 5-means clustering of the first 4 principal components.

## 4. Polygenic risk scores

In both cohorts, polygenic risk scores (PRS) were calculated using PRSice v.2 (Choi & O'Reilly, 2019) and genotyped variants. PRSice computes scores in an independent (target) sample by calculating the weighted sum of trait-associated alleles using summary data from GWAS discovery samples. SNPs in linkage disequilibrium (r2 ≥ 0.1 [250-kb window]) were removed. We used the default average option that calculates the ratio between the PRS and the number of alleles included in each individual; PRS were standardised (mean=0, SD=1). PRS were calculated at 11 p-value thresholds PT (5e-8, 1e-5, 1e-3, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 1) and the most predictive PT was selected. Logistic regression models were used to estimate associations between the phenotype and each PRS adjusting for covariates of six genetic ancestry principal components, assessment centre and batch effects in UKB (Supplementary code 3), and six genetic ancestry principal components and primary care practice in EXCEED. The proportion of variance explained by PRS on the liability scale was estimated according to Lee et al. (Lee et al., 2012), assuming MDD prevalence of 10.8% for case-control comparisons (Lim et al., 2018), and using the observed scale for case only comparisons. For case-control comparisons using different definitions of major depressive disorder (MDD) in UKB, for converting Nagelkerke $R^2$ to the liability scale we considered the relative frequency of each MDD phenotype compared to our primary care-defined MDD (at least two diagnostic codes for a

depressive disorder), and we multiplied it by the prevalence reported in the general population (10.8%).

**References**


Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562*(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*, 7. https://doi.org/10.1186/s13742-015-0047-8

Choi, S. W., & O'Reilly, P. F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, *8*(7). https://doi.org/10.1093/gigascience/giz082

Haro, J. M., Arbabzadeh-Bouchez, S., Brugha, T. S., De Girolamo, G., Guyer, M. E., Jin, R., Lepine, J. P., Mazzi, F., Reneses, B., Vilagut, G., Sampson, N. A., & Kessler, R. C. (2006). Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health Surveys. *International Journal of Methods in Psychiatric Research*, *15*(4), 167–180. https://doi.org/10.1002/mpr.196

Kessler, R. C., Andrews, G., Mroczek, D., Ustun, B., & Wittchen, H.-U. (1998). The World Health Organization Composite International Diagnostic Interview short-form (CIDI-SF). *International Journal of Methods in Psychiatric Research*, *7*(4), 171–185. https://doi.org/10.1002/mpr.47

Lee, S. H., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2012). A Better Coefficient of Determination for Genetic Profile Analysis: A Better Coefficient of Determination. *Genetic Epidemiology*, *36*(3), 214–224. https://doi.org/10.1002/gepi.21614

Lim, G. Y., Tam, W. W., Lu, Y., Ho, C. S., Zhang, M. W., & Ho, R. C. (2018). Prevalence of Depression in the Community from 30 Countries between 1994 and 2014. *Scientific Reports*, *8*(1), 2861. https://doi.org/10.1038/s41598-018-21243-x

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics (Oxford, England)*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Smith, D. J., Nicholl, B. I., Cullen, B., Martin, D., Ul-Haq, Z., Evans, J., Gill, J. M. R., Roberts, B., Gallacher, J., Mackay, D., Hotopf, M., Deary, I., Craddock, N., & Pell, J. P. (2013). Prevalence and characteristics of probable major depression and bipolar disorder within UK biobank: Cross-sectional study of 172,751 participants. *PloS One*, *8*(11), e75362. https://doi.org/10.1371/journal.pone.0075362

UK Biobank. (2019, September 1). *Primary Care Linked Data*. http://biobank.ndph.ox.ac.uk/showcase/showcase/docs/primary_care_data.pdf

Warren, H. R., Evangelou, E., Cabrera, C. P., Gao, H., Ren, M., Mifsud, B., Ntalla, I., Surendran, P., Liu, C., Cook, J. P., Kraja, A. T., Drenos, F., Loh, M., Verweij, N., Marten, J., Karaman, I., Lepe, M. P. S., O'Reilly, P. F., Knight, J., … UK Biobank CardioMetabolic Consortium BP working group. (2017). Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nature Genetics*, *49*(3), 403–415. https://doi.org/10.1038/ng.3768