The contributions of common and rare genetic variation to different measures of mood and anxiety disorder in UK Biobank

Ioanna Katzourou PhD[1], LINC Consortium[#], Inês Barroso PhD[2], Lauren Benger MSc[1], Andrés Ingason PhD[3], Daniel Stow PhD[4], Ruby Tsang PhD[5], Megan Wood PhD[6], George Kirov PhD[1], James Walters PhD[1], Michael J. Owen PhD[1,7], Peter Holmans PhD[1], Marianne B.M. van den Bree PhD[1,7]

ABSTRACT

Background: Mood and anxiety disorders co-occur and share symptoms, treatments and genetic risk, but it is unclear whether combining them into a single phenotype would better capture genetic variation. The contribution of common genetic variation to these disorders has been investigated using a range of measures, however the differences in their ability to capture variation remain unclear, while the impact of rare variation is mostly unexplored.

Aims: We aimed to explore the contributions of common genetic variation and Copy Number Variations associated with risk of psychiatric morbidity (P-CNVs) to different measures of internalising disorders.

Method: We investigated eight definitions of mood and anxiety disorder, and a combined internalising disorder, derived from self-report questionnaires, diagnostic assessments, and electronic healthcare data (EHR). Association of these definitions with polygenic risk scores (PRSs) of major depressive disorder and anxiety disorder, as well as presence of a P-CNV, was assessed.

Results: The effect sizes of both PRSs and P-CNVs were similar for mood and anxiety disorder. Compared to mood and anxiety disorder, internalising disorder resulted in higher prediction accuracy for PRSs, and increased significance of associations with P-CNVs for most definitions. Comparison across the eight definitions showed that PRSs had higher prediction accuracy and effect sizes for stricter definitions, whereas P-CNVs were more strongly associated with EHR- and self-report- based definitions.

Conclusions: Future studies may benefit from using a combined internalising disorder phenotype, and may need to consider that different phenotype definitions may be more informative depending on whether common or rare variation is studied.

INTRODUCTION

BACKGROUND

Mood and anxiety disorders are highly prevalent, affecting over 300 million people worldwide(1), have a detrimental impact on the quality of life of affected individuals and those close to them(2) and result in increased healthcare costs (3,4). Evidence from psychiatric genetics research indicates that currently used diagnostic boundaries do not accurately reflect the underlying shared genetic architecture of psychopathology(5). Research findings demonstrate that psychiatric conditions are highly polygenic, and that there is overlap in genetic risk between diagnostic groups(6).

Mood and anxiety disorders tend to co-occur(7), respond to similar pharmacological and psychological treatments(8,9), and have been indicated to involve similar neurobiological mechanisms(10). Despite extensive research, their aetiology is still incompletely understood, however it is evident that genetic predisposition plays a role, with common(11–14) as well as rare variants(15,16) having been associated with both disorders. There is abundant evidence that the two disorders share genetic liability(17–21). There may therefore be benefits to conducting analyses where the two disorders are combined into a single internalising disorder phenotype, which may be better able to capture genetic variation than each disorder individually, however to date there has been no published research examining this.

The contribution of common genetic variation to risk of psychiatric disorders is conferred by multiple variants, each of modest effect size, resulting in limited predictive power(22).

Polygenic risk scores (PRS) integrate the effect sizes of multiple variants throughout the genome and create quantifiable scores that are better suited for risk prediction than individual variants(23). In addition to common variants, a range of rare genetic variants, such as Copy Number Variants (CNVs), have been reported to have a large effect on an individual's risk of psychiatric outcomes(24). CNVs are rare sub-microscopic genomic re-arrangements, including deletions or duplications and a range of these have been found to greatly increase risk of neurodevelopmental (25) and psychiatric disorders (P-CNVs from here onwards). For example, these CNVs have been associated with elevated rates of anxiety disorders in youth(24) and anxiety and mood disorders in adulthood(15,16). While both P-CNVs and common variation are implicated in the development of anxiety and mood disorders, the majority of the literature focuses exclusively on either common or rare variants, with few studies exploring both(25).

Large case-control studies are required to detect the contributions of genetic variants to psychiatric disorders. The recruitment of patients in such studies is resource-intensive and can introduce biases inherent to the selection of participants (26). An alternative approach is utilising the breadth of phenotypic information that large-scale population-based biobanks can provide, where extensive sample sizes can contribute to the discovery of new genotype-phenotype associations. This is particularly the case for common disorders, as indicated by recent genome-wide association studies (GWAS) of Major Depressive Disorder (MDD) in cohorts of very large size(11,12). However, the findings of this type of studies crucially depend on the correct assignment of participants to case or control status, which will differ depending on the measures that have been selected out of those available in the biobank. It is likely that different phenotypic definitions differ in the ability to classify individuals into cases and controls, and therefore their ability to capture genetic variation. Definitions that are more

sensitive to capturing disease-specific genetic variation will be more informative in identifying the biological pathways involved in these conditions and ultimately in guiding future intervention strategies. In the case of MDD, it has been reported that genetic analyses conducted on measures involving minimal phenotyping (e.g. self-reported seeking of medical attention or diagnosis) result not only in a reduced single nucleotide polymorphism (SNP) heritability, but also in genetic associations that are less specific to MDD, showing greater overlap with other neuropsychiatric traits(27). In contrast, more strict, narrowly-defined phenotypic measures were found to yield greater SNP heritability and to better capture MDD-specific genetic variation(27). This work did, however, not include some other commonly used measures of mood disorder, such as primary care records or medication use, while it also remains unclear if similar findings apply to anxiety disorders. Furthermore, it is unknown if these results extend to risk attributable to rare genetic variation.

The great majority of the psychiatric genetics literature to date has focused on individuals of European ancestry, with few studies examining individuals of different genetic ancestries(28). This has meant that individuals of non-European ancestries have been removed from data sets before genetic analysis is undertaken. Novel analytical strategies are now providing opportunities to conduct genomic studies in ancestrally diverse and admixed populations, allowing for more inclusive and representative studies(29,30), increasing the generalisability of findings.

AIMS

The purpose of this study was to investigate the contributions of common genetic variation and P-CNVs to a range of self-reported and EHR-derived definitions of anxiety and mood disorders and to evaluate whether combining these disorders into an internalising disorder phenotype improves the ability to capture genetic influences. We focused on the UKBB cohort because it is a large population-based resource, combining information on anxiety and mood disorders from a range of different sources with genetic data.

Specifically, we aimed to:

1. Evaluate whether the predictive accuracy of PRSs for anxiety and MDD improves when information on anxiety and mood is combined into an internalising disorder phenotype, in comparison to analyses based on individual phenotypes of anxiety and mood disorder.
2. Investigate the differences in association with these PRSs between eight different definitions for each of mood, anxiety and internalising disorder, based on self-report questionnaires, diagnostic interview, medication use and primary care and hospital admission EHR data.
3. Investigate if the presence of a P-CNV is more strongly associated with a) the combined internalising disorder phenotype than with individual phenotypes of anxiety or mood disorder and b) any of the eight definitions for each of mood, anxiety and internalising disorder mentioned above in 2.

These analyses included participants in UKBB of all ancestries and PRS were adjusted for ancestral differences.

METHOD

PARTICIPANTS

The UKBB is a prospective study of over 500,000 individuals living in the United Kingdom (UK)(31). Participants aged between 40 and 69 years old were recruited between 2006 and 2010. They attended a baseline assessment as well as multiple repeat assessments. The UKBB received ethical approval from the North West - Haydock Research Ethics Committee (reference 16/NW/0274). Participants provided electronic signed consent at recruitment. This study was conducted under application number 79704.

PHENOTYPING

Four main sources of information relevant to anxiety and mood disorder were identified in UKBB. These were:

- A touchscreen questionnaire completed by participants during initial recruitment to the study at recruitment centers.
- A nurse-led interview completed at recruitment to which participants were invited if they stated in the touchscreen questionnaire that they had been diagnosed with certain long-term conditions or were currently taking medication.
- Linked electronic healthcare records (EHR), including hospital admission records (available for the whole cohort) and primary care records (available for ~40% of the cohort).

- The mental health questionnaire (MHQ), which was an online follow-up sent to all participants with a valid email address(7).

The numbers of individuals with available data for these four sources are summarized in Figure 1. Using these sources of information, eight ways of defining internalising disorder were derived, summarized in Figure 1. For each definition, individuals that were established to have either a mood or anxiety disorder or both were classified as having an internalising disorder.

[FIGURE 1]

Figure 1. Data sources for internalising disorder definitions in UKBB. Self-report (coded 1) was defined as having reported during the nurse-led interview a diagnosis of depression or post-natal depression for mood disorder and anxiety/panic attacks for anxiety disorder. Medication self-report (coded 2) was defined as having reported during the nurse-led interview currently being on a prescription of any antidepressant for mood disorder and any antidepressant and/or benzodiazepine apart from temazepam for anxiety disorder. Help-seeking behaviour (coded 3) was defined as having answered yes to either 'have you ever seen a GP for depression, tension or nerves?' or 'have you ever seen a psychiatrist for depression, tension or nerves?', thus help-seeking behaviour is identical for mood and anxiety disorders. Minimal phenotyping (coded 4 in Figure 1) was defined according to Smith et al. (2013)(32) for mood disorder and as having endorsed the help-seeking phenotype and in addition having a score of 10 or above on the generalised anxiety disorder 7 (GAD-7)(33) for anxiety disorder. CIDI-SF (coded 5) was defined using items of the MHQ that correspond to the Composite International Diagnostic Interview – Short Form (CIDI-SF)(34) diagnostic criteria for lifetime major depression for mood disorder and lifetime generalised anxiety disorder (GAD) for anxiety disorder. MHQ self-report (coded 6) was defined as having reported in the MHQ having had a diagnosis of depression for mood disorder or social anxiety or social phobia, agoraphobia, panic attacks, anxiety, nerves and GAD for anxiety disorder. The presence of mood and anxiety disorder in hospital admission records (coded 7) and primary care records (coded 8) was established using lists of clinical codes curated by the MULTIPLY(35) project and amended to exclude specific phobias and other non-specific codes (Supplementary Material).

Individuals who had a diagnosis of schizophrenia or bipolar disorder either recorded in EHR or self-reported at the nurse-led interview or the MHQ (N=4,214) were excluded from the analyses, as these disorders often share symptomatology with and could be misdiagnosed as internalising disorders. Thus, a sample of maximum n=496,412 was available for analysis.

GENETIC ANALYSES

GENETIC QUALITY CONTROL

The UKBB had imputed genotype data to the Haplotype Reference Consortium and the UK10K Consortium using the IMPUTE4 software(31). Preliminary quality control of the genotype data had also been performed by UKBB(31). We performed additional quality control using PLINK 2.0(36) to filter for variants with a low INFO score (< 0.9), high missingness (> 0.05), low minor allele frequency (< 0.01) or variants departing from Hardy-Weinberg equilibrium ($p < 10^{-6}$) and individuals with high missingness (> 0.05) or sex discordance. Sex chromosomes were excluded. Kinship estimates were computed to identify individuals related to the second degree (KING(37) $r^2 > 0.0884$) and one individual from each related pair was removed at random. After quality control, 449,646 individuals and 6,899,626 variants were retained.

POLYGENIC RISK SCORE (PRS) GENERATION

Anxiety disorder PRS was calculated using summary statistics from the iPSYCH anxiety disorder GWAS(38) (4,584 cases, 19,225 controls). MDD PRS was calculated using summary statistics from the latest Psychiatric Genomics Consortium (PGC) MDD GWAS(12), excluding the UKBB cohort (45,621 cases, 97,674 controls). PRS-CS(39) was used for PRS calculation. PRS-CS is a Bayesian algorithm that can infer posterior effect sizes of SNPs via continuous shrinkage(39),

therefore avoiding the need for linkage disequilibrium pruning and p-value thresholding. The inferred posterior effect sizes were used for PRS generation on PLINK 2.0(36).

POST-HOC PRS ADJUSTMENT FOR ANCESTRY

In order to produce PRSs that are on the same scale across individuals from different ancestries, we adjusted them for ancestral differences in mean and variance using the 1000Genomes dataset as reference, as described by Khan *et al.*(30). The UKBB and 1000Genomes(40) datasets were merged, retaining only variants present in both (N SNPs = 4,944,504). The two datasets were then pruned using PLINK(36) --indep-pairwise 500 50 0.05, resulting in 567,216 retained variants. FlashPCA(41) was used to generate principal components (PCs) in the 1000Genomes datasets and UKBB participants were projected onto these components. PRS were calculated for 1000Genomes as described above.

First, using the 1000Genomes dataset, the PRS was regressed against the first five PCs to generate coefficients and residuals. The coefficients and residual variance from these models were then used to produce ancestry adjusted PRS(30). The raw PRS was standardised by subtracting the mean of the predicted PRS and dividing by the residual variance. The same procedure was subsequently performed in UKBB using the PC projections. The distribution of the adjusted PRS between different ancestries was visually inspected in both 1000Genomes and UKBB. The post-hoc adjustment was performed using R(42).

PRS ASSOCIATION ANALYSIS

The association of each of the eight definitions of mood, anxiety and internalising disorder with the standardised PRS of anxiety disorder and MDD was tested using logistic regression, adjusting for sex, age and the first ten genetic principal components to account for population structure. The predictive accuracy of the PRS was estimated using the receiver operating characteristic area under curve (AUC), calculated by the pROC package in R(43). The proportion of phenotypic variance explained by the PRS was estimated by comparing Nagelkerke's pseudo-R^2 of the full model (PRS and covariates) with the null model (covariates only). The effect sizes of the PRSs from the logistic regression models were compared between the different definitions of internalising disorder. Due to the considerable sample overlap between the definitions, equations six and seven from Lin and Sullivan(44) were used to calculate the variance of the effect size difference, and thence a z-score for testing the significance of the difference. To assess the presence of sex-specific genetic effects, we also conducted the logistic regression analysis described above for each of the eight definitions of internalising disorder while adding an interaction term between sex and PRS. The analysis was performed using R(42).

CNV CALLING

The process of the calling of the CNVs in UKBB is described in detail elsewhere(45). Briefly, calling was performed using PennCNV-Affy 1.0.3 protocols(46). Samples were excluded if they carried 30 or more CNVs, had a waviness factor greater than 0.03 or less than –0.03, a single-nucleotide polymorphism call rate lower than 96%, or log R ratio SD higher than 0.35, while

CNVs were excluded if they were covered by fewer than 20 probes, had a density coverage of less than 1 probe per 20,000 base pairs, or a confidence score lower than 10, resulting in 454,254 individuals with available CNV call data.

CNV ASSOCIATION ANALYSIS

A set of 54 CNVs that have previously been associated with an increased risk of a psychiatric disorder (45) (P-CNVs) were studied. CNVs that were observed fewer than five times were excluded, resulting in 33 P-CNVs that were included in all subsequent analyses. First, the association of each of the eight definitions of mood, anxiety and internalising disorder with the presence of any of the P-CNVs was assessed using logistic regression, adjusting for age, sex and the first ten genetic principal components to account for population structure. Then, the association of the eight definitions of internalising disorders with each individual P-CNV was assessed in the same way. The effect sizes of the presence of a P-CNV from the logistic regression models were compared between the different definitions of internalising disorder as described above. We also conducted regression analysis for each of the eight definitions conditioning anxiety disorder on mood disorder. The purpose of this analysis was to test whether the associations of P-CNV with anxiety disorder were independent of those with mood disorder. Non-independence of these associations provides further rationale for combining mood and anxiety disorder into an internalising disorder phenotype, since this will increase power without losing associations specific to the rarer disorder (anxiety). Then, to assess the presence of sex-specific genetic effects, we also conducted the logistic regression analysis described above for each of the eight definitions of internalising disorder while adding an

interaction term between sex and the presence of a P-CNV. Finally, we assessed the association of each of the eight definitions of internalising disorder with the presence of each of the P-CNVs individually using logistic regression, adjusting for age, sex and the first ten genetic principal components to account for population structure. Statistical analyses were performed using R(42).

JOINT ANALYSIS OF PRS AND CNV

To examine if PRS and P-CNV act independently or whether there is evidence common and rare variation interact to increase the risk of internalising disorder, logistic regression analyses were performed as described previously, including the main effects of PRS, P-CNV and an interaction term PRS*P-CNV. The statistical analyses were performed using R(42).

RESULTS

SUMMARY STATISTICS

The prevalence of each definition of anxiety, mood and internalising disorder is shown in Figure 2 and Supplementary Table 1. The frequency of each definition of internalising disorder for males and females is shown in Supplementary Figure 1 and Supplementary Table 2. All definitions had a significantly higher prevalence in females compared to males. For the combined internalising disorder phenotype, the definition with the highest prevalence was help-seeking (33.7%, 169,330 cases). Because of the way this was queried in the touchscreen

questionnaire, this definition was the same for mood, anxiety and internalising disorder (see Figure 1). The lowest prevalence was found for initial self-report (7.74%, 31,869 cases). MHQ self-report and CIDI-SF had a high prevalence (29.57% and 24.71%, respectively) although the number of cases identified was not particularly large (46,514 and 38,887, respectively). For internalising disorder, the most common combinations of definitions are illustrated in Supplementary Figure 2 while the number of definitions present for each individual is shown in Supplementary Figure 3.

[FIGURE 2]

Figure 2. Prevalence of each definition of anxiety, mood and internalising disorder. For each definition, individuals with missing values were removed from the calculation.

Tetrachoric correlations were calculated for each pair of phenotypes, as presented in Supplementary Figure 4. All eight definitions of internalising disorder were significantly positively correlated.

PRS ANALYSIS

The distributions of the PRS pre- and post-ancestry adjustment in the reference populations of the 1000Genomes dataset and in the UKBB populations (based on self-report of ethnicity) were visually examined. The distributions were notably different prior to adjustment, whereas post adjustment the shapes of the distributions were similar (Supplementary Figures 5 and 6).

The association of the adjusted PRSs with the definitions of mood, anxiety and internalising disorder were assessed and the OR with 95% CI, p-value, Nagelkerke's pseudo R^2 and AUC were calculated (Supplementary Table 3 for mood disorder, Supplementary Table 4 for anxiety disorder and Table 1 for internalising disorder). All phenotypes were significantly associated with both MDD and anxiety disorder PRS after Bonferroni correction for multiple testing), with the MDD PRS showing a more significant association and larger effect sizes for all phenotypes compared to the anxiety disorder PRS. The AUC of the models was used to quantify the prediction accuracy of the PRS for the different phenotypes, as illustrated in Figure 3. Combining mood and anxiety disorder into internalising disorder resulted in increased AUC for both PRS for all definitions, with the exception of help-seeking behaviour (which, as explained In the Summary Statistics section is the only definition that does not distinguish between mood and anxiety disorder, and is therefore identical for all three conditions, including internalising disorder). For most definitions, the ORs of the association of each of the two PRSs with mood and anxiety disorder were similar (Supplementary Tables 3 and 4), while combined internalising disorder phenotype yielded a higher AUC than either mood or anxiety disorder separately. This indicates that combining the disorders results in increased prediction accuracy in PRS analyses.

[FIGURE 3]

Figure 3. Prediction accuracy of MDD (top) and anxiety disorder (bottom) PRS for the eight definitions of mood disorders, anxiety disorders and internalising disorders.

[TABLE1]

Table 1. Association metrics of the adjusted MDD and anxiety PRS with the eight internalising disorder phenotypes. (OR: odds ratio, AUC: area under the curve, R^2: Nagelkerke's pseudo R^2)

The effect sizes (ORs) for each of the PRSs did not differ significantly between the definitions. The highest predictive accuracy (AUC) for both PRSs was observed for the CIDI-SF definition. The AUC was highest for the MHQ-derived phenotypes (CIDI-SF and minimal phenotyping) while the AUC was lowest for EHR-derived phenotypes, with the AUC for help-seeking behaviour and the self-reported phenotypes being in between. The Nagelkerke's pseudo-R^2 was low, ranging between 0.50-0.88% for MDD PRS and 0.10-0.19% for anxiety PRS. Age and female sex were significantly positively associated with all the definitions. We found no significant interaction between sex and either of the PRSs.

Restricting the analyses to individuals of European ancestry (self-reported ethnicity white British, white Irish or any other white background, N = 418,120) gave similar association results to those of the full cohort (Supplementary Table 5). Restricting to individuals of non-European ancestry (all other self-reported ethnicities, N = 25,782, Supplementary Table 6) led to lower ORs than in the European ancestry sample, although the differences were not significant. Associations with PRS also yielded less significant p-values than in the analyses of European ancestry (Supplementary Table 7), with p-values not meeting the significance threshold for anxiety PRS (Bonferroni-corrected p-value threshold = 3.125×10^{-3}, N tests = 16) for many phenotypic definitions. Comparison of association of ethnicity-adjusted versus unadjusted PRS in individuals of non-European ancestry indicated that adjusted PRSs resulted in higher ORs, however the difference was not significant (Supplementary Table 7).

CNV ANALYSIS

The total number of individuals with a P-CNV was 7,454. The number of individuals with a P-CNV endorsing each of the definitions of mood, anxiety and internalising disorder are shown in Table 2. The associations between presence of any of the P-CNVs and the definitions of mood, anxiety and internalising disorder are shown in Table 2. For most of the definitions, the ORs were similar between mood, anxiety and internalising disorder, and using the combined internalising disorder definitions resulted in similar ORs and increased significance. None of the associations between P-CNV and anxiety disorder were significant after conditioning on mood disorder. This highlights the interdependency of the mood and anxiety disorder phenotypes and provides further rationale for combining them into an internalising disorder phenotype to increase power without losing associations specific to the rarer disorder (anxiety). Of note, the only negative association in Table 2, between the presence of a P-CNV and the minimal definition for anxiety disorder was no longer significant after controlling for mood disorder.

[TABLE 2]

Table 2. Results of the logistic regression of P-CNV carrier status with the eight definitions of mood, anxiety and internalising disorder and number of individuals with a P-CNV and each definition of mood disorder, anxiety disorder and internalising disorder

When correcting for multiple testing (Bonferroni-corrected p-value threshold 2.083×10^{-3}, N tests = 24), the presence of a P-CNV was significantly associated with six of the definitions of internalising disorders, but not with the CIDI-SF or minimal phenotyping. The highest effect sizes were observed for EHR-derived and self-reported definitions (initial self-report and medication self-report). The OR for help-seeking behaviour was significantly lower than those for all other definitions that had a significant association with the presence of a P-CNV, while

the OR for hospital admission records was significantly higher than those for primary care records and MHQ self-report, but not initial self-report and medication self-report, and there was no significant difference between the ORs for MHQ self -report, primary care records, initial self-report and medication self-report. Minimal and CIDI-SF were not significantly associated with the presence of a P-CNV, therefore the ORs for these definitions were not included in the comparisons. The p-values of the pairwise OR comparisons are given in Supplementary Table 8. We found no significant interaction between sex and the presence of a P-CNV for any of the outcomes.

We subsequently examined the association of the 38 P-CNVs with the definitions of internalising disorders individually. The number of individuals with each of the 38 P-CNVs are shown in Supplementary Table 9. The results are illustrated in Supplementary Figure 7. The pattern of association of individual P-CNVs with the eight definitions of internalising disorders was complex, with no individual P-CNV showing association with all definitions. For some of the individual P-CNVs the pattern of association was similar to that of the aggregated P-CNVs, for example 15q11.2 duplication and 16p13.11 deletion were significantly positively associated with EHR-derived and self-reported definitions and not with MHQ- and questionnaire-based definitions. 17p13.3 duplication and 22q11.2 deletion, on the other hand, were positively associated with CIDI-SF and MHQ self-report, while 22q11.2 distal deletion and 15q24 duplication were negatively associated with CIDI-SF and MHQ self-report, findings that were in contrast to the aggregated P-CNV results. No P-CNVs was associated with Help-seeking behaviour.

PRS AND CNV INTERACTION ANALYSIS

No evidence of significant interaction of either PRS with presence of a P-CNV was found for any of the definitions of internalising disorders (Bonferroni-corrected p-value threshold 3.125×10^{-3}, N tests =16). The results are shown in Supplementary Table 10.

DISCUSSION

Our study aimed to explore the genetic architecture of internalising disorders and to assess the genetic burden associated with different definitions of these disorders derived from a number of different types of assessments. We hypothesised that mood and anxiety disorder can be grouped into a combined internalising disorder phenotype, based on previous work that has illustrated that the two disorders correlate phenotypically and genetically(19–21). We constructed eight different phenotypic definitions of mood, anxiety and internalising disorder, aiming to determine if there are ways of defining the disorder that better capture genetic liability. We aimed to examine the effect of both common and rare genetic variation across the genome and included UKBB participants of all ethnic backgrounds. We found that combining mood and anxiety disorder into an internalising disorder resulted in a higher predictive accuracy in PRS analyses, regardless of the way in which the phenotype data were obtained. For P-CNVs, we found that combining the disorders resulted in similar or higher effect sizes and stronger associations for some of the definitions. Moreover, we found that stricter definitions of internalising disorders resulted in better prediction accuracy in PRS analyses, while EHR-derived and self-reported definitions had the highest effect sizes in analysis of P-CNVs.

We combined information on mood and anxiety disorder into a single internalising disorder and compared the association of PRS derived from GWAS of MDD(12) and anxiety disorder(38) on this phenotype with those for mood and anxiety disorder measured individually. Anxiety and mood disorder had similar associations with each of the two PRS. The combined internalising disorder phenotype resulted in similar or higher effect sizes, more significant associations and higher predictive accuracy than the mood and anxiety disorder definitions individually. This was the case across all eight disorder definitions. While the increased significance could result from the higher number of affected individuals for internalising disorder, the higher AUCs would indicate that the strengthening of the results also stems from the genetic overlap between anxiety and mood disorder. Anxiety disorder PRS had lower OR and AUC than MDD PRS across all eight definitions, even when predicting anxiety disorder. However, the GWAS used to derive the anxiety PRS had a smaller sample size(12), and thus the anxiety PRS is likely to be less powerful than the MDD PRS. The AUCs we found ranged from 0.75 to 0.63, similar to ones reported in literature for depression PRS (0.57)(12), bipolar disorder PRS (0.65)(47), schizophrenia PRS (0.72)(48), and Alzheimer's disease PRS (0.69)(49).

When comparing the association of the PRSs with the eight definitions of internalising disorder, both PRSs had a higher prediction accuracy for MHQ-derived definitions of internalising disorders that include standardized questionnaires, such as CIDI-SF and minimal phenotyping. These definitions are totally or partially based on parts of the MHQ(7). EHR-derived definitions, like primary care records and hospital admissions showed the lowest prediction accuracy. This is in agreement with earlier investigations of the genetic liability of different definitions of depression in UKBB that have found that depression diagnosed using CIDI-SF has the highest SNP heritability and help-seeking behaviour the lowest(50,51). Cai *et al*. (2020)(50) found that

461 the genetic liability of minimally defined depression, a phenotype similar to the help-seeking
462 behaviour used in this study, is less specific to depression and includes more liability shared
463 with other psychiatric traits. In the same study, PRS derived from help-seeking behaviour had
464 the highest prediction accuracy for depression in a separate sample. However, when deriving
465 PRS from each definition using the same sample size, a CIDI-based definition of depression
466 resulted in the highest AUC(50). When we derived PRS of MDD and tested its association with
467 the different definitions, we also found the highest prediction accuracy was achieved when
468 using CIDI-SF definitions. We have, therefore, shown that conclusions regarding the genetic
469 liability of different definitions of depression also extend to definitions of anxiety and
470 combined internalising disorder. Additionally, our study included a wider range of phenotypic
471 definitions that are often used in public health studies, including primary care records and
472 medication use. Interestingly, we found that primary care records, hospital admission records
473 and self-reported medication use had the lowest AUC for both MDD and anxiety PRS, indicating
474 that they capture common genetic variation less well than the other definitions we studied.
475
476 We compared UKBB participants with at least one of 38 CNVs previously associated with high
477 risk of a psychiatric condition (45) with those without these P-CNVs. We assessed the
478 association of the presence of these P-CNVs with the different definitions of mood, anxiety and
479 internalising disorder. The ORs for anxiety and mood disorder were similar, and combining the
480 disorders resulted in more significant associations for some of the definitions as it increased
481 the statistical power due to an increase in the number of affected individuals. Three of the
482 definitions (initial self-report, CIDI-SF and minimal) were endorsed by fewer individuals with a
483 P-CNV for anxiety disorder than for mood disorder, therefore for these the association with
484 internalising disorder was likely mostly driven by mood disorder.

We found the presence of a P-CNV had the highest effect size for EHR-derived definitions (primary care and hospital admission records) and definitions based on the self-report at recruitment (initial self-report and medication self-report). CIDI-SF and minimal phenotyping were not associated with the presence of a P-CNV. This is in direct contrast to the results of the PRS analyses. A possible explanation could be that internalising disorders caused by common genetic variation are less severe, and therefore less likely to result in the use of healthcare services compared to internalising disorders that are associated with a rare genetic variant. On the other hand, individuals with a P-CNV have been found to have a higher risk of developing physical and mental health multimorbidity(52,53), it is therefore likely that they have more contact with health services than individuals without these CNVs. This could mean that any evidence of internalising disorder is also more likely to be queried and diagnosed by a physician and treated, and therefore recorded in their EHR or self-reported as a diagnosis or a medication. Finally, the presence of an interactive effect between PRS and the presence of a P-CNV was explored. There was no significant interaction between the presence of a P-CNV and either of the PRSs for any of the definitions of internalising disorder, which is in agreement with recent findings of the effect of common and rare variation on psychopathology in UKBB(25). This suggests that the risk conferred by common and rare genetic variants is independent, at least for the definitions of internalising disorder examined. Interestingly, we found no significant interaction between sex and any of the genetic risk factors we examined, which indicates that these do not act in a sex-specific way.

While the majority of studies into the genetics of mental health conditions have been restricted to a single population, usually comprising of individuals of European ancestry(28),

we aimed to include the whole UKBB cohort in our study and not restrict our analysis to a single ethnic background, as including ethnically diverse populations in genetic studies can uncover differing biological risk factors and aid in combatting health inequalities. As PRS derived from European cohorts have been found to perform sub-optimally in populations of non-European ancestry(54), we attempted to adjust the PRS for ancestry effects using an ancestrally diverse dataset as our reference. Prior to adjustment, there were considerable differences in the mean and variance of both MDD and anxiety disorder PRS, post-adjustment however the PRSs distributions were notably more aligned, particularly so for the MDD PRS. While the adjustment did not result in perfect alignment of the distributions, this method allowed for including UKBB participants of all ancestries in the analysis.

There are limitations to this study. Firstly, the MDD GWAS that was used for PRS generation in this analysis was based on a larger sample size(12) and thus better captured the genetic risk than the anxiety disorder GWAS(38). However, the two disorders are genetically and phenotypically correlated(6,19), and our results indicate that the MDD PRS also captures genetic risk for anxiety disorders. Moreover, while UKBB is one of the largest population cohorts with genetic data available worldwide and contains rich phenotypic information, it has been found to be impacted by selection bias, with participants having better health and higher socioeconomic status than the general population in the UK(55). Additionally, UKBB was designed as a prospective population cohort study of middle and older age(31), recruiting participants between 40 and 69 years of age, which makes it susceptible to survival bias. Internalising disorders are associated with premature mortality(56), therefore it is likely that individuals with the most severe manifestations of these disorders would not be included in such a cohort. Finally, UKBB participants who have completed the MHQ have been found to

be of higher socioeconomic status and better overall health than the average UKBB participant, with a further bias towards individuals of European descent(7). This is particularly important for the CNV analyses, as the number of individuals with a P-CNV that completed the MHQ was low (1,950, 26.16% of individuals with a P-CNV, compared to 31.61% completion for individuals without a P-CNV), therefore the sample size might not have been sufficient to uncover significant associations with these phenotypes.

In conclusion, this study aimed to explore the genetic architecture of internalising disorder definitions. Our results indicate that combining mood and anxiety disorders into an internalising disorder phenotype can be of benefit in genetic analyses looking at both common and rare variants. The optimal definition of internalising disorders for use in genetic studies depends on the type of genetic researchers aim to uncover. While more clinically robust definitions of internalising disorders, like the CIDI-SF diagnostic criteria, seem preferable when examining common variation, using EHR- or self-report-based definitions might be the optimal choice when rare variation is of interest.

AUTHOR DETAILS

[1]Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, UK

[2]Medical School, University of Exeter, UK

[3]Institute of Biological Psychiatry, Capital Region of Denmark, Denmark

[4]Wolfson Institute for Population Health, Queen Mary University of London, UK

[5]Bristol Medical School, University of Bristol, UK

[6]School of Psychology, University of Leeds, UK

[7]Neuroscience and Mental Health Innovation Institute, Division of Psychological Medicine and Clinical Neurosciences, Cardiff University, UK

[#]A list of authors and their affiliations appears at the end of the paper.

Corresponding author: Marianne van den Bree

Email: vandenbreemb@cardiff.ac.uk

Phone: +44 (0)29 2068 8433

Fax: +44(0)29 2068 7068

Address: Cardiff University Centre for Neuropsychiatric Genetics and Genomics, 2nd Floor Hadyn Ellis Building, Cardiff University, Maindy Road, Cathays, Cardiff, United Kingdom, CF24 4HQ

ACKNOWLEDGEMENTS

This research has been conducted using the UK Biobank Resource under Application Number 79704.

Full list of LINC members:

Marianne B. M. van den Bree, George Kirov, Michael J. Owen, James T. R. Walters, Peter A. Holmans, Jane Lynch, Ioanna K. Katzourou, Lowri O'Donovan (Cardiff University, UK)

David A. van Heel, Sarah Finer, Daniel Stow (Queen Mary University of London, UK)

Golam M. Khandaker, Nicholas J. Timpson, John A. A. MacLeod, Julie P. Clayton, Ruby S. M. Tsang, Jane Sprackman, Shahid Khan (University of Bristol, UK)

Inês Barroso, Rupert A. Payne (University of Exeter, UK)

Mark Mon-Williams, Megan L. Wood (University of Leeds, UK)

Hilary C. Martin (Wellcome Sanger Institute, UK)

Thomas Werge, Andrés Ingason (Institute of Biological Psychiatry, Denmark)

We thank Dr Jack Underwood for comments on the manuscript.

We thank the members of the LINC study public advisory group for their contribution.

FUNDING

This work was funded by the Tackling Multimorbidity at Scale Strategic Priorities Fund programme (MR/W014416/1) delivered by the Medical Research Council and the National

Institute for Health Research in partnership with the Economic and Social Research Council and in collaboration with the Engineering and Physical Sciences Research Council.

AUTHOR CONTRIBUTION

Study conceptualisation and design: IKK, MvdB, GK, MJO, PH, JW, LINC. Analytical consultation and interpretation: IKK, MvdB, GK, MJO, PH, JW, AI, RT, DS, IB. UKBB data curation: IKK. Genetic data preparation: IKK. Supervision: MvdB, GK, MJO, PH, JW. All listed authors critically edited the manuscript.

DATA AVAILABILITY

Relevant data is available from UK Biobank subject to standard procedures (www.ukbiobank.ac.uk).

DECLARATION OF INTEREST

Michael J Owen is receiving research grants from Takeda Pharmaceuticals and Akrivia Health. The remaining authors declare no competing interests.

Supplementary information is available on the BJPsych Open website.

REFERENCES

1. World Health Organization. Depression and Other Common Mental Disorders—Global Health Estimates [Internet]. World Health Organization; 2017. Available from: https://iris.who.int/bitstream/handle/10665/254610/W?sequence=1

2. Hohls JK, König HH, Quirke E, Hajek A. Anxiety, Depression and Quality of Life—A Systematic Review of Evidence from Longitudinal Observational Studies. Int J Environ Res Public Health. 2021 Nov 16;18(22):12022.

3. König H, König HH, Konnopka A. The excess costs of depression: a systematic review and meta-analysis. Epidemiol Psychiatr Sci. 2019 Apr 5;29:e30.

4. Konnopka A, König H. Economic Burden of Anxiety Disorders: A Systematic Review and Meta-Analysis. PharmacoEconomics. 2020 Jan;38(1):25–37.

5. Smoller JW, Andreassen OA, Edenberg HJ, Faraone SV, Glatt SJ, Kendler KS. Psychiatric Genetics and the Structure of Psychopathology. Mol Psychiatry. 2019 Mar;24(3):409–20.

6. The Brainstorm Consortium. Analysis of shared heritability in common disorders of the brain. Science. 2018 Jun 22;360(6395):eaap8757.

7. Davis KAS, Coleman JRI, Adams M, Allen N, Breen G, Cullen B, et al. Mental health in UK Biobank – development, implementation and results from an online questionnaire completed by 157 366 participants: a reanalysis. BJPsych Open. 2020 Mar;6(2):e18.

8. Goodwin GM, Stein DJ. Generalised Anxiety Disorder and Depression: Contemporary Treatment Approaches. Adv Ther. 2021;38(Suppl 2):45–51.

9. Garber J, Brunwasser SM, Zerr AA, Schwartz KTG, Sova K, Weersing VR. Treatment and Prevention of Depression and Anxiety in Youth: Test of Cross-Over Effects. Depress Anxiety. 2016 Oct;33(10):939–59.

10. Oathes DJ, Patenaude B, Schatzberg AF, Etkin A. Neurobiological signatures of anxiety and depression in resting-state functional magnetic resonance imaging. Biol Psychiatry. 2015 Feb 15;77(4):385–93.

11. Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019 Mar;22(3):343–52.

12. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018 May;50(5):668–81.

13. Otowa T, Hek K, Lee M, Byrne EM, Mirza SS, Nivard MG, et al. Meta-analysis of genome-wide association studies of anxiety disorders. Mol Psychiatry. 2016 Oct;21(10):1391–9.

14. Levey DF, Gelernter J, Polimanti R, Zhou H, Cheng Z, Aslan M, et al. Reproducible Genetic Risk Loci for Anxiety: Results From ~200,000 Participants in the Million Veteran Program. Am J Psychiatry. 2020 Mar 1;177(3):223–32.

15. Kendall KM, Rees E, Bracher-Smith M, Legge S, Riglin L, Zammit S, et al. Association of Rare Copy Number Variants With Risk of Depression. JAMA Psychiatry. 2019 Aug;76(8):818–25.

16. Adams RL, Baird A, Smith J, Williams N, van den Bree MBM, Linden DEJ, et al. Psychopathology in adults with copy number variants. Psychol Med. 53(7):3142–9.

17. Kendler KS, Myers J. The boundaries of the internalizing and externalizing genetic spectra in men and women. Psychol Med. 2014 Feb;44(3):10.1017/S0033291713000585.

18. Mather L, Blom V, Bergström G, Svedberg P. An Underlying Common Factor, Influenced by Genetics and Unique Environment, Explains the Covariation Between Major Depressive Disorder, Generalized Anxiety Disorder, and Burnout: A Swedish Twin Study. Twin Res Hum Genet Off J Int Soc Twin Stud. 2016 Dec;19(6):619–27.

19. Thorp JG, Campos AI, Grotzinger AD, Gerring ZF, An J, Ong JS, et al. Symptom-level modelling unravels the shared genetic architecture of anxiety and depression. Nat Hum Behav. 2021 Oct;5(10):1432–42.

20. Waszczuk MA, Zavos HMS, Gregory AM, Eley TC. The phenotypic and genetic structure of depression and anxiety disorder symptoms in childhood, adolescence, and young adulthood. JAMA Psychiatry. 2014 Aug;71(8):905–16.

21. Morneau-Vaillancourt G, Coleman JRI, Purves KL, Cheesman R, Rayner C, Breen G, et al. The genetic and environmental hierarchical structure of anxiety and depression in the UK Biobank. Depress Anxiety. 2020 Jun;37(6):512–20.

22. Gratten J, Wray NR, Keller MC, Visscher PM. Large-scale genomics unveils the genetic architecture of psychiatric disorders. Nat Neurosci. 2014 Jun;17(6):782–90.

23. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007 Oct;17(10):1520–8.

24. Chawner SJRA, Owen MJ, Holmans P, Raymond FL, Skuse D, Hall J, et al. Genotype–phenotype associations in children with copy number variants associated with high neuropsychiatric risk in the UK (IMAGINE-ID): a case-control cohort study. Lancet Psychiatry. 2019 Jun 1;6(6):493–505.

25. Mollon J, Schultz LM, Huguet G, Knowles EEM, Mathias SR, Rodrigue A, et al. Impact of Copy Number Variants and Polygenic Risk Scores on Psychopathology in the UK Biobank. Biol Psychiatry. 2023 Oct 1;94(7):591–600.

26. Palumbo SA, Robishaw JD, Krasnoff J, Hennekens CH. Different Biases in Meta-Analyses of Case-Control and Cohort Studies: An Example from Genomics and Precision Medicine. Ann Epidemiol. 2021 Jun;58:38–41.

27. Cai N, Revez JA, Adams MJ, Andlauer TFM, Breen G, Byrne EM, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet. 2020 Apr;52(4):437–47.

28. Brown JEH, Young JL, Martinez-Martin N. Psychiatric genomics, mental health equity, and intersectionality: A framework for research and practice. Front Psychiatry. 2022 Dec 21;13:1061705.

29. Meng X, Giannakopoulou O, Navoly G, Levey D, Mitchell B, Oliveira AM, et al. 17. ANCESTRY-AWARE MIXED MODEL GWAS OF MAJOR DEPRESSION CHARTS A PATH FOR INCLUSIVE AND DIVERSE GENETIC RESEARCH. Eur Neuropsychopharmacol. 2023 Oct 1;75:S65–6.

30. Khan A, Turchin MC, Patki A, Srinivasasainagendra V, Shang N, Nadukuru R, et al. Genome-wide polygenic score to predict chronic kidney disease across ancestries. Nat Med. 2022 Jul;28(7):1412–20.

31. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.

32. Smith DJ, Nicholl BI, Cullen B, Martin D, Ul-Haq Z, Evans 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. 2013;8(11):e75362.

33. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006 May 22;166(10):1092–7.

34. Kessler RC, Andrews G, Mroczek D, Ustun B, Wittchen HU. The World Health Organization Composite International Diagnostic Interview short-form (CIDI-SF). Int J Methods Psychiatr Res. 1998;7(4):171–85.

35. Eto F, Samuel F, Finer S. MULTIPLY-Initiative: Version 1.1 [Internet]. 2023. Available from: doi:10.5281/zenodo.7643566.

36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

37. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010 Nov 15;26(22):2867–73.

38. Meier SM, Trontti K, Purves KL, Als TD, Grove J, Laine M, et al. Genetic Variants Associated With Anxiety and Stress-Related Disorders: A Genome-Wide Association Study and Mouse-Model Study. JAMA Psychiatry. 2019 Sep 1;76(9):924–32.

39. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019 Apr 16;10(1):1776.

40. Siva N. 1000 Genomes project. Nat Biotechnol. 2008 Mar 1;26(3):256–256.

41. Abraham G, Inouye M. Fast Principal Component Analysis of Large-Scale Genome-Wide Data. PLoS ONE. 2014 Apr 9;9(4):e93766.

42. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2022.

43. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011 Mar 17;12(1):77.

44. Lin DY, Sullivan PF. Meta-analysis of genome-wide association studies with overlapping subjects. Am J Hum Genet. 2009 Dec;85(6):862–72.

45. Kendall KM, Rees E, Escott-Price V, Einon M, Thomas R, Hewitt J, et al. Cognitive Performance Among Carriers of Pathogenic Copy Number Variants: Analysis of 152,000 UK Biobank Subjects. Biol Psychiatry. 2017 Jul 15;82(2):103–10.

46. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, et al. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007 Jan 11;17(11):1665–74.

47. Mullins N, Forstner AJ, O'Connell KS, Coombes B, Coleman JRI, Qiao Z, et al. Genome-wide association study of over 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat Genet. 2021 Jun;53(6):817–29.

48. Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature. 2022 Apr;604(7906):502–8.

49. Leonenko G, Baker E, Stevenson-Hoare J, Sierksma A, Fiers M, Williams J, et al. Identifying individuals with high risk of Alzheimer's disease using polygenic risk scores. Nat Commun. 2021 Jul 23;12(1):4506.

50. Cai N, Revez JA, Adams MJ, Andlauer TFM, Breen G, Byrne EM, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet. 2020 Apr;52(4):437–47.

51. Glanville KP, Coleman JRI, Howard DM, Pain O, Hanscombe KB, Jermy B, et al. Multiple measures of depression to enhance validity of major depressive disorder in the UK Biobank. BJPsych Open. 2021 Feb 5;7(2):e44.

52. Crawford K, Bracher-Smith M, Owen D, Kendall KM, Rees E, Pardiñas AF, et al. Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. J Med Genet. 2019 Mar;56(3):131–8.

53. Finucane B, Oetjens MT, Johns A, Myers SM, Fisher C, Habegger L, et al. Medical manifestations and health care utilization among adult MyCode participants with neurodevelopmental psychiatric copy number variants. Genet Med Off J Am Coll Med Genet. 2022 Mar;24(3):703–11.

54. Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, et al. Principles and methods for transferring polygenic risk scores across global populations. Nat Rev Genet. 2023 Aug 24;

55. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol. 2017 Nov 1;186(9):1026–34.

56. Plana-Ripoll O, Pedersen CB, Agerbo E, Holtz Y, Erlangsen A, Canudas-Romo V, et al. A comprehensive analysis of mortality-related health metrics associated with mental disorders: a nationwide, register-based cohort study. The Lancet. 2019 Nov 16;394(10211):1827–35.