Biological Psychiatry

Archival Report

# Polygenic Analyses Show Important Differences Between Major Depressive Disorder Symptoms Measured Using Various Instruments

Lianyun Huang, Sonja Tang, Jolien Rietkerk, Vivek Appadurai, Morten Dybdahl Krebs, Andrew J. Schork, Thomas Werge, Verena Zuber, Kenneth Kendler, and Na Cai

## ABSTRACT

**BACKGROUND:** Symptoms of major depressive disorder (MDD) are commonly assessed using self-rating instruments like the Patient Health Questionnaire-9 (PHQ-9) (current symptoms) and the Composite International Diagnostic Interview Short-Form (CIDI-SF) (worst-episode symptoms). We performed a systematic comparison between them for their genetic architecture and utility in investigating MDD heterogeneity.

**METHODS:** Using data from the UK Biobank ($n$ = 41,948–109,417), we assessed the single nucleotide polymorphism heritability and genetic correlation ($r_g$) of both sets of MDD symptoms. We further compared their $r_g$ with non-MDD traits and used Mendelian randomization to assess whether either set of symptoms has more genetic sharing with non-MDD traits. We also assessed how specific each set of symptoms is to MDD using the metric polygenic risk score pleiotropy. Finally, we used genomic structural equation modeling to identify factors that explain the genetic covariance between each set of symptoms.

**RESULTS:** Corresponding symptoms reported through the PHQ-9 and CIDI-SF have low to moderate genetic correlations ($r_g$ = 0.43–0.87), and this cannot be fully attributed to different severity thresholds or the use of a skip structure in the CIDI-SF. Both Mendelian randomization and polygenic risk score pleiotropy analyses showed that PHQ-9 symptoms are more associated with traits that reflect general dysphoria, whereas the skip structure in the CIDI-SF allows for the identification of heterogeneity among likely MDD cases. Finally, the 2 sets of symptoms showed different factor structures in genomic structural equation modeling, reflective of their genetic differences.

**CONCLUSIONS:** MDD symptoms assessed using the PHQ-9 and CIDI-SF are not interchangeable; the former better indexes general dysphoria, while the latter is more informative about within-MDD heterogeneity.

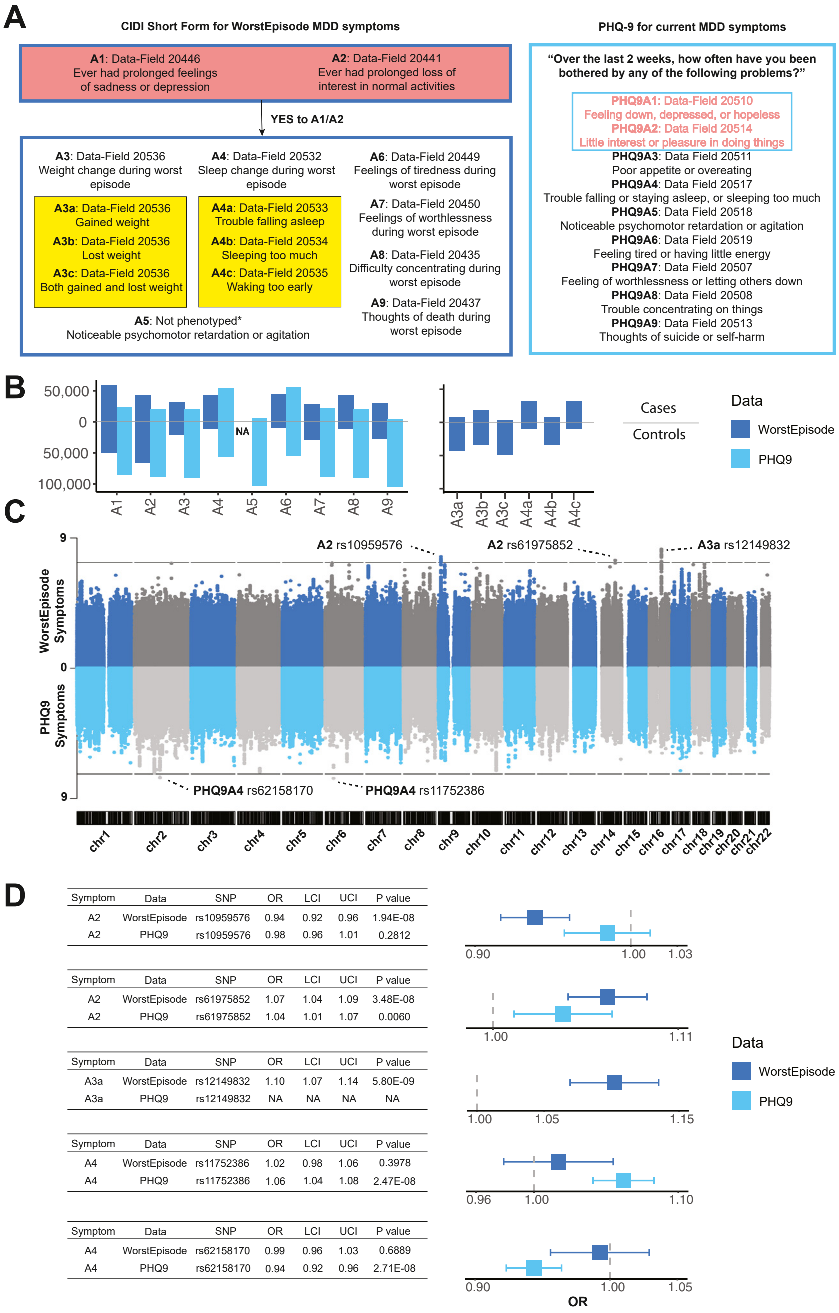
https://doi.org/10.1016/j.biopsych.2023.11.021

Two sets of symptom-level data on major depressive disorder (MDD) are available in the UK Biobank (1) through the self-administered online Mental Health Questionnaire (MHQ) (2). First, current symptoms of MDD are assessed through the Patient Health Questionnaire-9 (PHQ-9) (3), a screening tool that scores the occurrence of all 9 DSM-5–based symptoms for MDD (4) in the past 2 weeks. A high score is used as an indicator of potential MDD and is the basis for recommending clinical assessment. Second, MDD symptoms experienced during the lifetime worst episode of MDD are assessed through the Composite International Diagnostic Interview Short-Form (CIDI-SF) (5,6), which contains a skip structure: 7 of 9 MDD symptoms are assessed only when 2 weeks of sad mood or anhedonia are reported (cardinal symptoms of MDD) (Figure 1A). Therefore, the CIDI-SF assesses only worst-episode symptoms in a clinically enriched population that has a smaller sample size (~50,000) (Figure 1B) than the PHQ-9 symptoms assessed in the general population (~100,000).

Previous studies have performed genome-wide association studies (GWASs) on individual items from the PHQ-9 (7). Likewise, genetic covariance–based factor analyses using genomic structural equation modeling (genomic SEM) (8) have been performed on PHQ-9 symptoms to identify symptom dimensions of MDD and how they overlap with symptoms of anxiety and neuroticism (7,9). In contrast, there are no genetic studies that have analyzed lifetime worst-episode symptoms or compared them with PHQ-9 symptoms despite the much wider use of the latter in factor analyses (10–12).

Previous findings suggest that symptoms in clinically enriched populations have different structures and meanings from symptoms that are measured in a general population (10). In this paper, we ask whether PHQ-9 symptoms capture the same underlying biology as worst-episode symptoms using data collected through the MHQ in the UK Biobank (1). We found that while PHQ-9 and worst-episode symptoms have similar liability scale single nucleotide polymorphism heritabilities ($h^2_{SNP}$), they have distinct genetic components. PHQ-9 symptoms have greater genetic sharing with subjective well-being, insomnia, neuroticism, anxiety, and exposure to stressful life events. Polygenic predictions on MDD and 50

ISSN: 0006-3223



| Symptom | Data | SNP | OR | LCI | UCI | P value |
|---|---|---|---|---|---|---|
| A2 | WorstEpisode | rs10959576 | 0.94 | 0.92 | 0.96 | 1.94E-08 |
| A2 | PHQ9 | rs10959576 | 0.98 | 0.96 | 1.01 | 0.2812 |

| Symptom | Data | SNP | OR | LCI | UCI | P value |
|---|---|---|---|---|---|---|
| A2 | WorstEpisode | rs61975852 | 1.07 | 1.04 | 1.09 | 3.48E-08 |
| A2 | PHQ9 | rs61975852 | 1.04 | 1.01 | 1.07 | 0.0060 |

| Symptom | Data | SNP | OR | LCI | UCI | P value |
|---|---|---|---|---|---|---|
| A3a | WorstEpisode | rs12149832 | 1.10 | 1.07 | 1.14 | 5.80E-09 |
| A3a | PHQ9 | rs12149832 | NA | NA | NA | NA |

| Symptom | Data | SNP | OR | LCI | UCI | P value |
|---|---|---|---|---|---|---|
| A4 | WorstEpisode | rs11752386 | 1.02 | 0.98 | 1.06 | 0.3978 |
| A4 | PHQ9 | rs11752386 | 1.06 | 1.04 | 1.08 | 2.47E-08 |

| Symptom | Data | SNP | OR | LCI | UCI | P value |
|---|---|---|---|---|---|---|
| A4 | WorstEpisode | rs62158170 | 0.99 | 0.96 | 1.03 | 0.6889 |
| A4 | PHQ9 | rs62158170 | 0.94 | 0.92 | 0.96 | 2.71E-08 |

**Figure 1.** Definition, sample size, and genome-wide association studies of worst-episode (Composite International Diagnostic Interview [CIDI]) and Patient Health Questionnaire-9 (PHQ-9) symptoms of major depressive disorder (MDD) in the UK Biobank. **(A)** Definitions of worst-episode and PHQ-9 symptoms of MDD in the UK Biobank. **(B)** Sample sizes for worst-episode and PHQ-9 symptoms. There is no data for worst-episode A5; subgroups of symptoms A3 and A4 are not assessed in the PHQ-9. **(C)** Miami plot for 14 worst-episode symptoms on top and 9 PHQ-9 symptoms at the bottom. Associations with *p* values smaller than 5 × $10^{-8}$ are considered as genome-wide significant and are indicated in the plot. **(D)** Forest plots and accompanying data showing the odds ratios (ORs) and *p* values at significant loci; statistics at the corresponding worst-episode or PHQ-9 symptoms are shown for comparison; error bars show 95% CIs of the OR estimates. LCI, lower 95% CI; SNP, single nucleotide polymorphism; UCI, upper 95% CI.

non-MDD phenotypes further showed that the skip structure in the CIDI-SF that involves making conditional assessments of noncardinal symptoms also ensures that they capture genetic sources of heterogeneity among likely MDD cases. Finally, factor analyses on genetic covariance from both sets of symptoms using genomic SEM identified different structures; high genetic correlations between PHQ-9 symptoms make most factor structures indistinguishable and therefore do not offer insights into MDD heterogeneity.

Therefore, we concluded that PHQ-9 and worst-episode symptoms do not reflect the same biology and cannot be used interchangeably in investigations into MDD heterogeneity; the former indexes genetic liability to general dysphoria in addition to MDD, while the latter captures genetic heterogeneity among likely MDD cases.

## METHODS AND MATERIALS

### Definition of Worst-Episode and PHQ-9 MDD Symptoms in the UK Biobank

Individual-level MDD symptom data are available for UK Biobank participants who answered the questions for MDD symptoms on the CIDI-SF and PHQ-9 conducted through an online mental health follow-up survey (MHQ, data category 138). Definitions of each worst-episode and PHQ-9 symptom, as well as their variations, can be found in the Supplemental Methods in Supplement 1 and Tables S1 and S2 in Supplement 2. To investigate PHQ-9 and worst-episode symptoms at the same GWAS power, we performed downsampling: for each corresponding pair of PHQ-9 and worst-episode symptoms, we downsampled the one with higher effective sample sizes ($N_{\mathrm{eff}}$) accounting for imbalance between cases and controls [$N_{\mathrm{eff}} = 4/(1/N_{\mathrm{cases}} + 1/N_{\mathrm{controls}})$] to the same $N_{\mathrm{eff}}$ of the one with lower $N_{\mathrm{eff}}$ (Tables S1 and S2 in Supplement 2), keeping its prevalence unchanged.

## Definition of Other Phenotypes in the UK Biobank

We selected other phenotypes in the UK Biobank, including insomnia, measures of subjective well-being, neuroticism, individual neuroticism items, anxiety symptoms, and stressful life event exposures, to test for genetic sharing with PHQ-9 and worst-episode symptoms. For all data fields in the UK Biobank, endorsement criteria, and sample sizes, see Table S4 in Supplement 2.

## Genome-wide Associations in the UK Biobank

Genome-wide association analysis was performed using imputed genotype data at 5,776,313 SNPs (minor allele frequency [MAF] $\geq$ 0.05, INFO score $\geq$ 0.9) in PLINK2 (13). We used 20 principal components (PCs) computed with flashPCA (14) on 337,198 White British individuals in the UK Biobank and genotyping arrays as covariates using a logistic regression model for binary phenotypes and a linear regression model for continuous phenotypes.

## GWAS on MDD from PGC29 and iPSYCH

We used the publicly available PGC29 GWAS summary statistics from the Psychiatric Genomics Consortium (15) (https://www.med.unc.edu/pgc/download-results). For iPSYCH, we performed GWAS using logistic regression in PLINK2 (13) on MDD defined by at least 1 specialty psychiatric care contact registered in the Danish Psychiatric Central Research Register (16) or the Danish National Patient Register (17) for an ICD-10 code of F32 or F33 in 2 independent iPSYCH cohorts, iPSYCH2012 (18) and iPSYCH2015i (19), with 42,250 and 23,351 unrelated individuals with European genetic ancestry, respectively. We used the top 10 genomic PCs from individuals in iPSYCH2012 and iPSYCH2015i computed using flashPCA (14) as covariates to control for population structure in each of the cohorts. Details of the iPSYCH cohorts can be found in the Supplemental Methods in Supplement 1.

## SNP Heritability and Genetic Correlation

To test for the heritability of each symptom and the genetic correlation ($r_g$) between pairs of symptoms, we performed linkage disequilibrium (LD) score regression implemented in LDSC version 1.0.1 (20), using in-sample LD scores estimated from 10,000 random White British UK Biobank (1) individuals at SNPs with MAF $>$ 0.05 as reference. We assumed that the population prevalence of each symptom was equal to its sample prevalence in the UK Biobank and estimated $h^2_{\mathrm{SNP}}$ on the liability scale for each symptom and $r_g$ between pairs of symptoms. One-sided paired $t$ tests were conducted on $r_g$s between the 2 sets of symptoms and non-MDD phenotypes in R.

## Univariable Mendelian Randomization

Two-sample univariable Mendelian randomization (UVMR) was performed using MendelianRandomization version 0.6.0 (21) implemented in R version 4.0.3. For each pair of exposures and outcomes, we used SNPs that were significantly associated ($p < 5 \times 10^{-6}$) genome-wide with each exposure as instruments. Clumping and LD pruning were performed with default settings with R library *ieugwasr*: clump_kb = 10,000, clump_r2 = 0.001. We tested the validity of instruments used in MR using the $F$ statistic (Tables S5, S6, and S9 in Supplement 2). Multiple testing corrections on the number of symptoms were performed separately for the PHQ-9 and worst-episode symptoms. To assess horizontal pleiotropy (22), we also conducted pleiotropy-robust weighted median MR (23), MR Egger (24), and Causal Analysis Using Summary Effect estimates (CAUSE) (25) to compare the MR estimates between different MR models (Supplemental Methods in Supplement 1, Tables S5, S6, and S9 in Supplement 2). One-sided paired $t$ tests were conducted on effect sizes from UVMR analyses in R to test whether effects for PHQ-9 or worst-episode symptoms were bigger (as either outcomes or exposures) (Tables S7 and S10 in Supplement 2).

## Multivariable MR Bayesian Model Averaging

We used MR Bayesian model averaging (26), a statistical learning algorithm for 2-sample multivariable MR, to select likely causal exposures from a larger set of candidate exposures. We selected independent genetic variants associated with any of the symptoms as instrumental variables for the multivariable MR model. We assumed that half of the tested items were expected causal risk factors (prior = 0.5) when iterating through all possible combinations of candidate models in the model averaging algorithm. We ranked symptoms according to their marginal inclusion probability and calculated the respective empirical $p$ values. Finally, we adjusted for multiple testing using the Benjamini-Hochberg false discovery rate method.

## Polygenic Risk Score and Polygenic Risk Score Pleiotropy

For all in-sample polygenic risk score (PRS) predictions in the UK Biobank, we obtained 10-fold cross-validation PRS on all PHQ-9 and worst-episode symptoms in the UK Biobank by performing a GWAS on each symptom 10 times, each time using 90% of the individuals, and building PRSs from these GWAS results with PRSice version 2 (27). We evaluated predictive accuracy for observed lifetime MDD and 50 non-MDD phenotypes (Table S4 in Supplement 2) in the held-out 10% for all 10 folds. For all PRS predictions in the UK Biobank phenotypes, we used 20 genomic PCs and the genotyping array used as covariates. We performed out-of-sample predictions of MDD diagnostic code (ICD-10 code of F32 or F33) in iPSYCH2012 and iPSYCH2015i using PRSs built from the same 10-fold GWAS on symptoms in the UK Biobank, as described in the Supplemental Methods in Supplement 1. For prediction of all binary phenotypes, we evaluated accuracy using Nagelkerke's $R^2$. For all quantitative phenotypes, we evaluated accuracy using ordinary $R^2$. PRS pleiotropy (28) was calculated for each PHQ-9 and worst-episode symptom using the ratio of its PRS predictions on 50 non-MDD phenotypes and its prediction on lifetime MDD in the UK Biobank or ICD-10–based MDD in iPSYCH cohorts (PRS pleiotropy = $R^2_{\text{non-MDD}}/R^2_{\text{MDD}}$).

### Factor Analysis Using Genomic SEM

Exploratory factor analysis was conducted using the *psych* library in R with the minimum residual (minres) extraction approach and "promax" rotation enabled by the GPArotation on PHQ-9 symptoms and worst-episode symptoms respectively, using genetic covariance matrices estimated with LD score regression. Solutions were assessed for their variance explained (retaining factors explaining $> 0.1$ of total variance), loadings onto individual symptoms (retaining factor loadings $> 0.2$), and match with previous findings. Confirmatory factor analysis (CFA) was then performed to assess model fit with the following metrics: Akaike information criterion, comparative fit index, and standardized root-mean-squared residual (retaining factor loadings $>$ 0.2). All analyses were performed under the framework of genomic SEM (8).

## RESULTS

### GWAS on PHQ-9 and Worst-Episode Symptoms

First, we performed GWASs on 14 worst-episode symptoms (Figure 1A; Table S1 in Supplement 2) and 9 PHQ-9 symptoms (Figure 1A; Table S2 in Supplement 2) in the UK Biobank. We found a total of 3 significantly associated loci (at $p < 5 \times 10^{-8}$) in 2 of the 14 worst-episode symptoms: A2 (anhedonia) and A3a (increase in appetite or weight) (Figure 1C; Table S3 in Supplement 2) and 2 significantly associated loci in 1 of 9 PHQ-9 symptoms: A4 (change in sleep) (Figure 1C; Table S3 in Supplement 2). None were significant after we corrected for multiple testing on the number of symptoms analyzed.

We asked whether each significant locus for worst-episode symptoms had a similar effect on the corresponding PHQ-9 symptom and vice versa. We found that only 1 of the SNPs significantly associated with worst-episode A2 had a significant association ($p < .05/5$) in the same direction of effect as A2 assessed through the PHQ-9 (rs61975852, $p = .006$) (Figure 1D; Table S3 in Supplement 2). Because GWAS power is different between the 2 sets of symptoms due to differences in sample sizes (Figure 1B), we calculated the effective sample size ($N_{\text{eff}}$) of each corresponding pair of PHQ-9 and worst-episode symptoms and downsampled the larger of the 2 to the lower $N_{\text{eff}}$, keeping prevalence constant (Tables S1 and S2 in Supplement 2). We found that although none of the significant loci remained significant in the GWAS after downsampling, their relative effect sizes in PHQ-9 and worst-episode symptoms remained the same (Figure S1 in Supplement 1).

### Genetic Differences Between PHQ-9 and Worst-Episode Symptoms

Next, we estimated liability scale SNP heritability ($h^2_{\text{SNP}}$) for all worst-episode and PHQ-9 symptoms using LD score regression. We did not find significant differences in $h^2_{\text{SNP}}$ estimates between the corresponding PHQ-9 and worst-episode symptoms (Figure 2A), but $r_g$s (29) between corresponding symptoms were mostly significantly different from unity, with the exception of A3 (change in appetite) and A9 (suicidal ideation) (Figure 2B). This shows that worst-episode symptoms and PHQ-9 symptoms are driven by partly distinct genetic factors.

Because worst-episode symptoms were assessed with a skip structure in the CIDI-SF and the PHQ-9 symptoms were not, we asked whether their low $r_g$ is due to this inherent difference in the instruments. To do this, we implemented the skip structure on PHQ-9 symptoms, obtaining "PHQ-9 Skip" symptoms (Supplemental Methods in Supplement 1; Table S2 in Supplement 2). We found that their $r_g$s with PHQ-9 symptoms were not significantly different from unity (Figure 2C), whereas their $r_g$s with worst-episode symptoms remained low (Figure 2D). In other words, low $r_g$ between PHQ-9 and worst-episode symptoms cannot be largely attributed to the skip structure in the CIDI-SF. Furthermore, we found that the low $r_g$s could not be fully accounted for by the different severity thresholds in the 2 sets of symptoms (Figure 2E, F; Supplemental Methods in Supplement 1).

### Reporting of PHQ-9 Symptoms Is Likely Due to General Dysphoria

It is well known that longstanding conditions that cause general dysphoria can lead to inflation of self-ratings of current symptoms with the PHQ-9 (30,31). To test this, we first asked whether 4 traits, including insomnia and measures of subjective well-being, had greater genetic sharing with reporting of the most similarly phrased PHQ-9 items than the corresponding worst-episode symptoms on the CIDI-SF (Table S4 in Supplement 2). We found that all 4 traits had higher $r_g$s with PHQ-9 symptoms than worst-episode symptoms, even though their error bars overlapped (1-sided paired $t$ test $p = .02$) (Figure 3A). Then we performed the same analysis with 4 other sets of phenotypes: generalized anxiety disorder (GAD) symptoms from the Generalized Anxiety Disorder 7 (GAD7) questionnaire in the MHQ (32,33), experience of stressful life events both recently and during one's lifetime (34–36), and individual neuroticism items from the Eysenck Personality Questionnaire Revised-Short Form (37) (Methods and Materials; Table S4 in Supplement 2). PHQ-9 symptoms had higher $r_g$s with all 4 sets of phenotypes than worst-episode symptoms (Figure 3B) (1-sided paired $t$ test $p < 2.2 \times 10^{-6}$). Overall, we found that PHQ-9 symptoms had greater genetic sharing with these non-MDD phenotypes that index general dysphoria.

Then we asked whether PHQ-9 and worst-episode symptom endorsement may be partly due to general dysphoria using MR, which assesses the association of genetic predictors of an exposure with an outcome (38,39). Using the inverse variance-weighted model in UVMR (Methods and Materials), we found that genetic effects on insomnia and

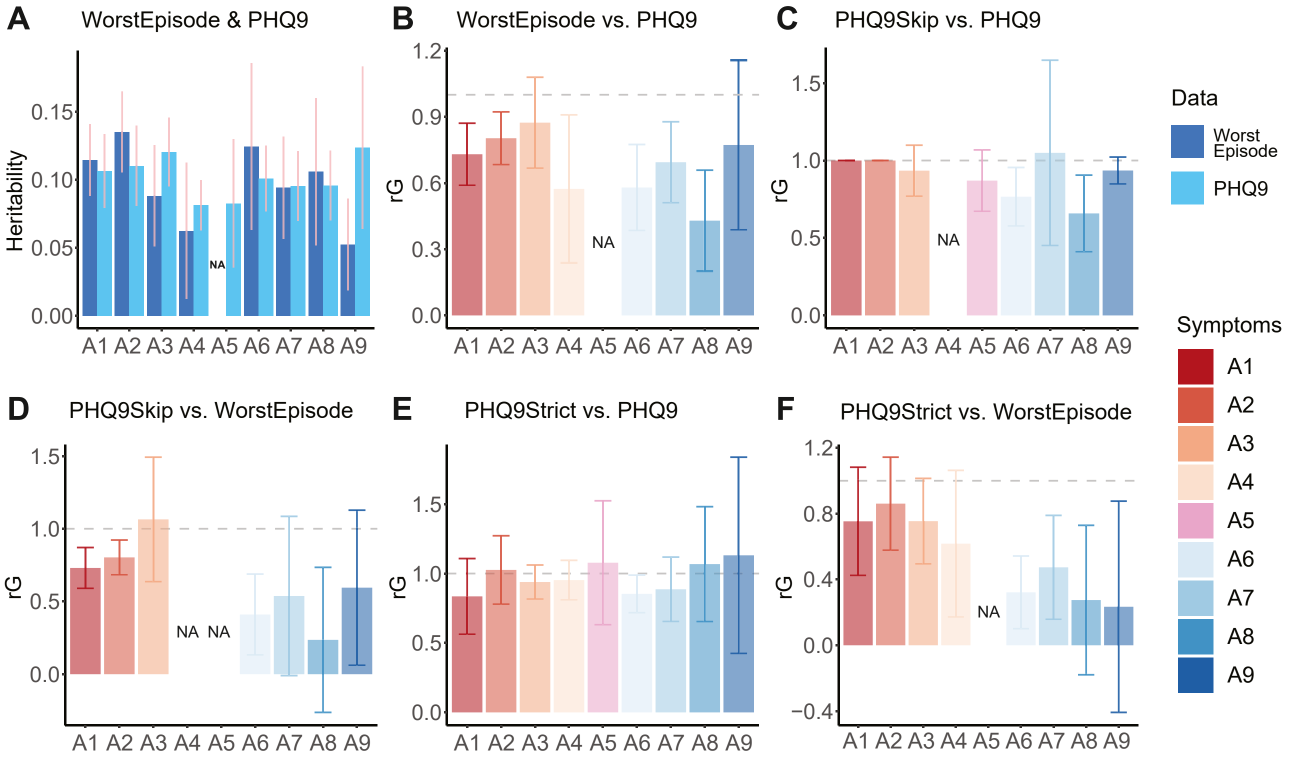


**Figure 2.** Single nucleotide polymorphism heritability ($h^2_{\text{SNP}}$) and genetic correlation ($r_g$) estimates from linkage disequilibrium score regression. **(A)** Liability scale $h^2_{\text{SNP}}$ estimates for each worst-episode and Patient Health Questionnaire-9 (PHQ-9) symptom, calculated assuming that their observed prevalences in the UK Biobank are equal to their population prevalences. **(B)** Genetic correlation between worst-episode and PHQ-9 symptoms. **(C)** Genetic correlation between PHQ-9 and PHQ-9 Skip symptoms. **(D)** Genetic correlation between PHQ-9 Skip and worst-episode symptoms. **(E)** Genetic correlation between PHQ-9 Strict and PHQ-9 symptoms. **(F)** Genetic correlation between PHQ-9 Strict and worst-episode symptoms. For all plots, horizontal dashed lines show $r_g$ = 1; error bars indicate 95% CIs of estimates; lack of convergence on the estimate or missing data is indicated as "NA." NA, not available.

measures of subjective well-being are associated with individual PHQ-9 symptoms with higher odds ratios than worst-episode symptoms (Figure 4A; Figures S2 and S5 in Supplement 1). The same is true for the neuroticism score and GAD (Figure 4C, D), but not lifetime trauma or recent stress (Figures S3, S6, and S7 in Supplement 1). These results remained when both sets of symptoms were downsampled to the same $N_{\text{eff}}$ (Figures S3, S6, and S7 in Supplement 1).

We verified these results with CAUSE, a MR method that accounts for horizontal pleiotropy, to distinguish causal effects from horizontal pleiotropy. Interestingly, the significant differences between neuroticism and GAD genetic effects on PHQ-9 and worst-episode symptoms are gone in CAUSE, demonstrating that neuroticism and GAD items are more likely pleiotropic with PHQ-9 symptoms than causal for them (Figure 4C, D; Tables S6 and S7 in Supplement 2). Furthermore, most UVMR results became insignificant when we used a multivariable MR approach based on Bayesian model averaging (Figure 4B) to assess whether the genetic effects on neuroticism and GAD are associated with the reporting of specific symptoms independently (Figure 4C, D; Figures S3 and S8 in Supplement 1). In other words, the contribution of measures of general dysphoria to both sets of MDD symptoms was not specific.

Finally, we asked whether episodic MDD leads to the endorsement of either set of symptoms and vice versa. To do this, we performed 2-sample UVMR (using the inverse variance weighted model and CAUSE) and MR Bayesian model averaging on either set of symptoms assessed in the UK Biobank with MDD assessed in external cohorts (PGC29, iPSYCH2012, and iPSYCH2015i). Overall, we found that genetic associations with worst-episode symptoms had greater effects on MDD, while genetic effects on MDD had greater effects on PHQ-9 symptoms (Supplemental Methods in Supplement 1; Figure 4E; Figures S4, S9–S11 in Supplement 1).

## Skip Structure Accounts for PRS Pleiotropy Difference Between Worst-Episode and PHQ-9 Symptoms

We further asked whether genetic studies on PHQ-9 symptoms are less likely to lead us to identify MDD-specific biology, by obtaining PRS pleiotropy of PHQ-9 and worst-episode symptoms for 50 non-MDD phenotypes in the UK Biobank (PRS pleiotropy = $R^2_{\text{non-MDD}}/R^2_{\text{MDD}}$) (28). A higher PRS pleiotropy means lower specificity of a PRS for MDD. For the MDD phenotype, we used LifetimeMDD (40) in the UK Biobank (Figure 5; Table S12 in Supplement 2) as well as ICD-10–based

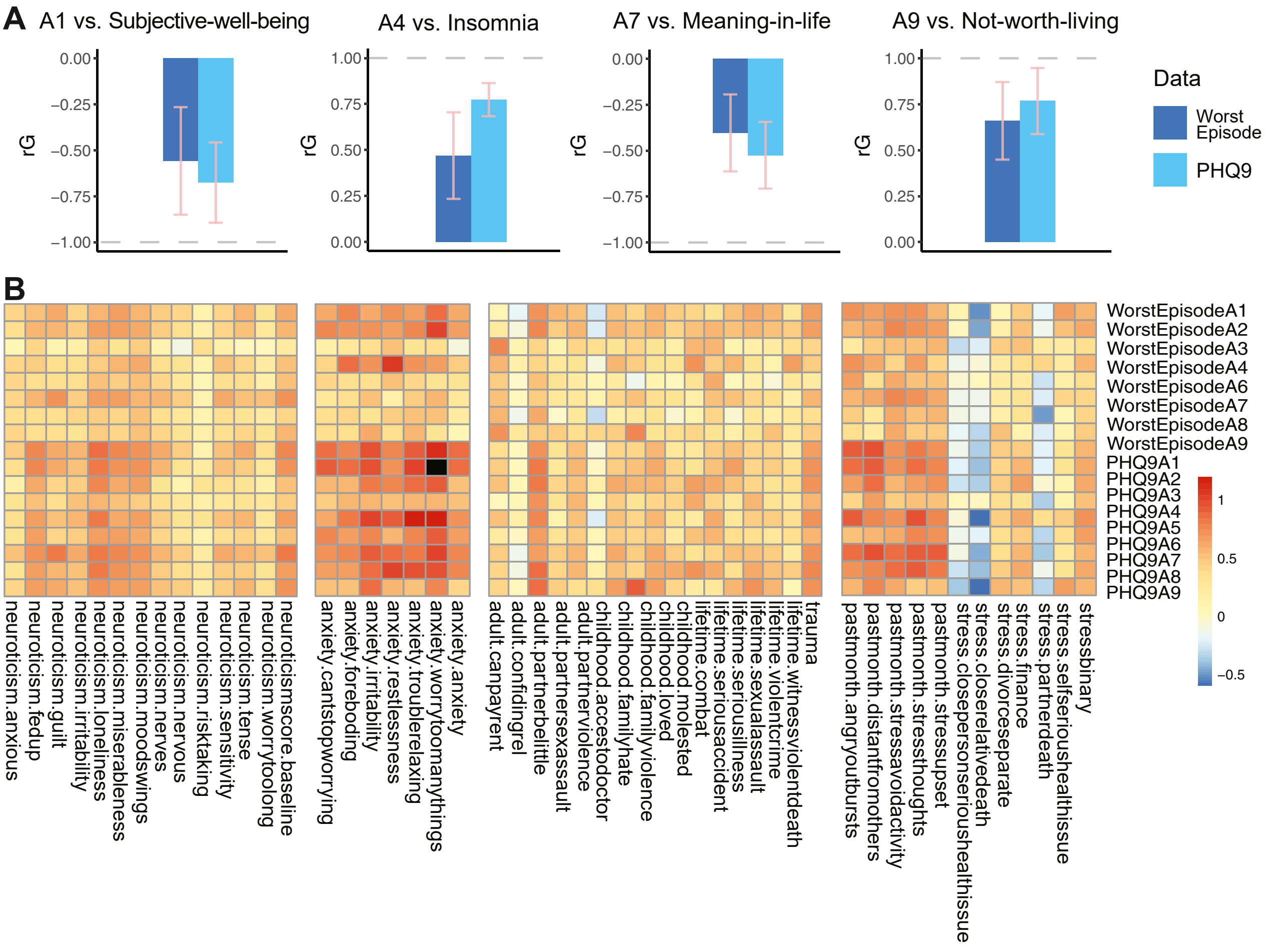


**Figure 3.** Genetic correlation ($r_g$) estimates between Patient Health Questionnaire-9 (PHQ-9) or worst-episode symptoms and non-major depressive disorder phenotypes. **(A)** Genetic correlation between depressed mood (A1) vs. subjective well-being; change of sleep (A4) vs. insomnia; feeling of worthlessness (A7) vs. finding meaning in life; and suicidal ideation (A9) vs. finding life not worth living in both worst-episode and PHQ-9 symptom definitions. Error bars indicate 95% CIs of the estimates. Bar directions indicate either positive or negative genetic correlation. **(B)** Genetic correlation between PHQ-9 or worst-episode symptoms of major depressive disorder vs. neuroticism items from the Eysenck Personality Questionnaire Revised-Short Form (37); anxiety symptoms from the Generalized Anxiety Disorder 7 questionnaire in the Mental Health Questionnaire (32,33); and experience of stressful life events both recently and during one's lifetime (34–36); black square indicates a lack of convergence on the estimate in linkage disequilibrium score regression.

MDD in iPSYCH2012 and iPSYCH2015i (Figures S5, S6, S13, and S14 in Supplement 1).

Overall, we found that worst-episode cardinal symptoms A1 (sad mood) and A2 (anhedonia) showed lower PRS pleiotropy across all examined non-MDD phenotypes than their PHQ-9 counterpart (Figure 5A, B; Figures S5 and S6 in Supplement 1), but the noncardinal worst-episode symptoms (A3–A9) showed much higher PRS pleiotropy than the corresponding PHQ-9 symptoms (Figure 5A, B; Figures S5 and S6 in Supplement 1). This apparently contradicts most of our results from the MR analyses, where we showed that PHQ-9 symptoms were more associated with genetic effects on non-MDD phenotypes. We hypothesize that this difference must come from the skip structure in the CIDI-SF which conditions worst-episode symptoms A3 to A9 on endorsement of symptoms A1 or A2, thereby removing the genetic liability in MDD that is shared with symptoms A1 and A2 and making them less predictive of MDD than their PHQ-9 counterparts. Consistent with this, we found that while worst-episode items A1 and A2 had higher $r_g$s with MDD than the corresponding PHQ-9 symptoms, the reverse was generally true for symptoms A3 to A9 (Figure S7 in Supplement 1).

To verify our hypothesis, we computed PRS pleiotropy of PHQ-9 Skip symptoms A3 to A9 and found that once the skip structure was applied, PHQ-9 Skip symptoms A3 to A9 were much less able to predict MDD, increasing their PRS pleiotropy for all non-MDD phenotypes (Figure 5A, B; Figures S5, S6; Supplemental Methods in Supplement 1). Furthermore, using genomic SEM, we found a high level of sharing between all PHQ-9 symptoms with MDD, which was greatly reduced when a skip structure was applied (Figure S8 in Supplement 1).

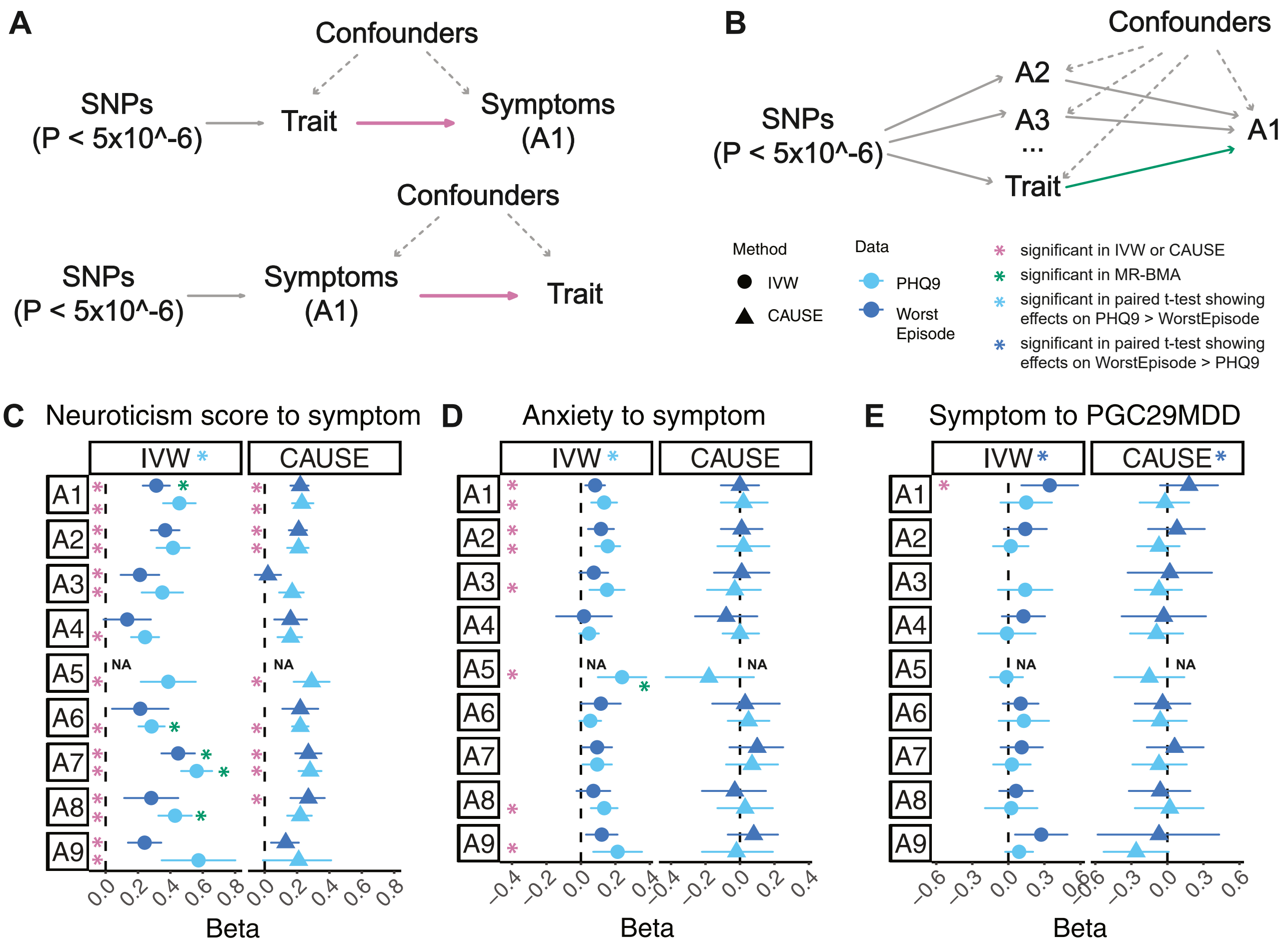


**Figure 4.** Mendelian randomization (MR) between major depressive disorder (MDD) symptoms and MDD/non-MDD phenotypes. **(A)** Directed acyclic graph demonstrating the univariable MR (UVMR) model in which MDD symptoms are either exposures or outcomes. **(B)** Directed acyclic graph demonstrating the MR Bayesian model averaging (MR-BMA) model. **(C, D)** Effect sizes in UVMR analyses under the inverse variance weighted (IVW) or Causal Analysis Using Summary Effect estimates (CAUSE) models, where neuroticism score or generalized anxiety disorder are exposures, and MDD symptoms are outcomes. Error bars show 95% CIs. **(E)** Effect sizes in UVMR analyses under the IVW or CAUSE models, where MDD symptoms are exposures, and MDD in PGC29 is the outcome. Error bars show 95% CIs. For plots **(C–E)**, pink asterisks indicate significance in UVMR tests for individual symptoms ($p < .05$ after Bonferroni correction); green asterisks indicate significance in MR-BMA for individual symptoms (with marginal inclusion threshold $> 0.5$ and $p < .05$ after false discovery rate correction); light blue asterisks indicate that effects on/of Patient Health Questionnaire-9 (PHQ-9) symptoms (as outcome/exposure) are larger than worst-episode symptoms in single-ended paired *t* tests ($p < .05$ after Bonferroni correction); dark blue asterisks indicate effects on/of worst-episode symptoms (as outcome/exposure) are larger than PHQ-9 symptoms in single-ended paired *t* tests ($p < .05$ after Bonferroni correction). PGC, Psychiatric Genomics Consortium; SNP, single nucleotide polymorphism.

## Worst-Episode Symptoms Are Better at Capturing Within-MDD Heterogeneity

Finally, we investigated the utility of worst-episode and PHQ-9 symptoms to identify genetically driven symptom dimensions of MDD with genomic SEM. We found that $r_g$ (and genetic covariance) (Supplemental Methods in Supplement 1) among PHQ-9 symptoms were significantly higher (mean $r_g$ = 0.80, SD = 0.14) (Figure 6A) than those among worst-episode symptoms (mean $r_g$ = 0.40, SD = 0.27) (Figure 6A), indicating that the worst-episode symptoms, especially A3 to A9, are more genetically heterogeneous than the PHQ-9 symptoms. As such, the choice between using either set of symptoms in factor analysis is therefore likely to make an important difference.

We first performed an exploratory factor analysis on the genetic covariance matrix of the 8 worst-episode symptoms (missing A5) and then ran confirmatory factor analyses (Supplemental Methods in Supplement 1) on the exploratory factor analysis solutions to assess model fits. We identified a 3-factor model as the best fit for worst-episode symptoms (Figure 6B; Table S15 in Supplement 2; Supplemental Methods in Supplement 1): the first "mood" factor loaded onto A2, A1, and A7 (all symptoms are ordered by loadings), the second "neurovegetative" factor loaded onto A3 and A4, and the third "psychomotor/cognitive" factor loaded onto A6, A8, A4, A9, and A7. This matches previous factor analyses based on phenotypic covariances (12,41,42) (Supplemental Methods in Supplement 1), although they are not always consistent with each other (43).

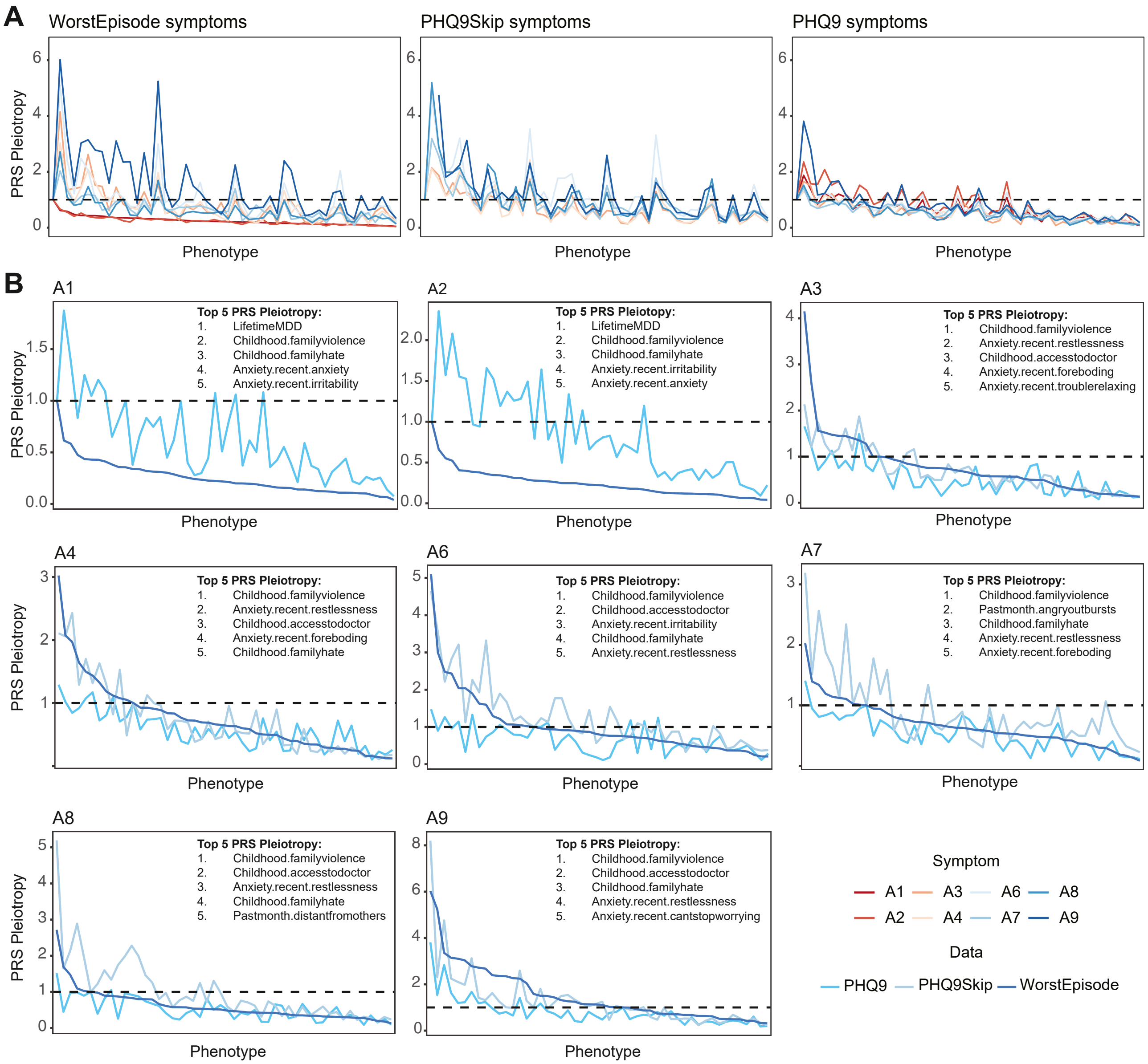


**Figure 5.** Polygenic risk score (PRS) pleiotropy of Patient Health Questionnaire-9 (PHQ-9) and worst-episode symptoms. **(A)** Mean PRS pleiotropy for major depressive disorder (MDD) across 10-fold cross-validation of worst-episode, PHQ-9, and PHQ-9 Skip symptoms on 50 phenotypes (including lifetime MDD and 50 non-MDD phenotypes, PRS pleiotropy = $R^2_{\text{non-MDD}}/R^2_{\text{MDD}}$); the MDD phenotype here is lifetime MDD (40) as defined in the UK Biobank; phenotypes on the x-axis across all 3 panels are ordered by worst-episode symptom A1 PRS pleiotropy in descending order. **(B)** Mean PRS pleiotropy for each symptom across 10-fold cross-validation; phenotypes on the x-axis in each panel are ordered by the worst-episode symptom PRS pleiotropy in descending order; the top 5 phenotypes in terms of PRS pleiotropy are indicated.

Then we performed the same analyses on the genetic covariance matrix of the PHQ-9 symptoms (Figure 6C; Table S15 in Supplement 2; Supplemental Methods in Supplement 1). In the best-fitting 2-factor solution identified from CFA, the first “psychological” factor loaded onto A2, A8, A1, A3, A7, and A9, and the second “somatic” factor loaded onto A5, A6, and A4 (Figure 6C). This structure is inconsistent with the 3-factor structure derived from the worst-episode symptoms as well as previous findings of 2-factor structures among MDD symptoms (10,44). Notably, the $r_g$ between factors 1 and 2 was much higher (0.90, SE = 0.04) than the average $r_g$ between factors observed in worst-episode symptoms (0.53, SE = 0.04), reflective of the high $r_g$ between all PHQ-9 symptoms. Consistent with this, removing A5 from PHQ-9 produced a 1-factor solution (Figure 6D). The $r_g$ between factors from both sets of

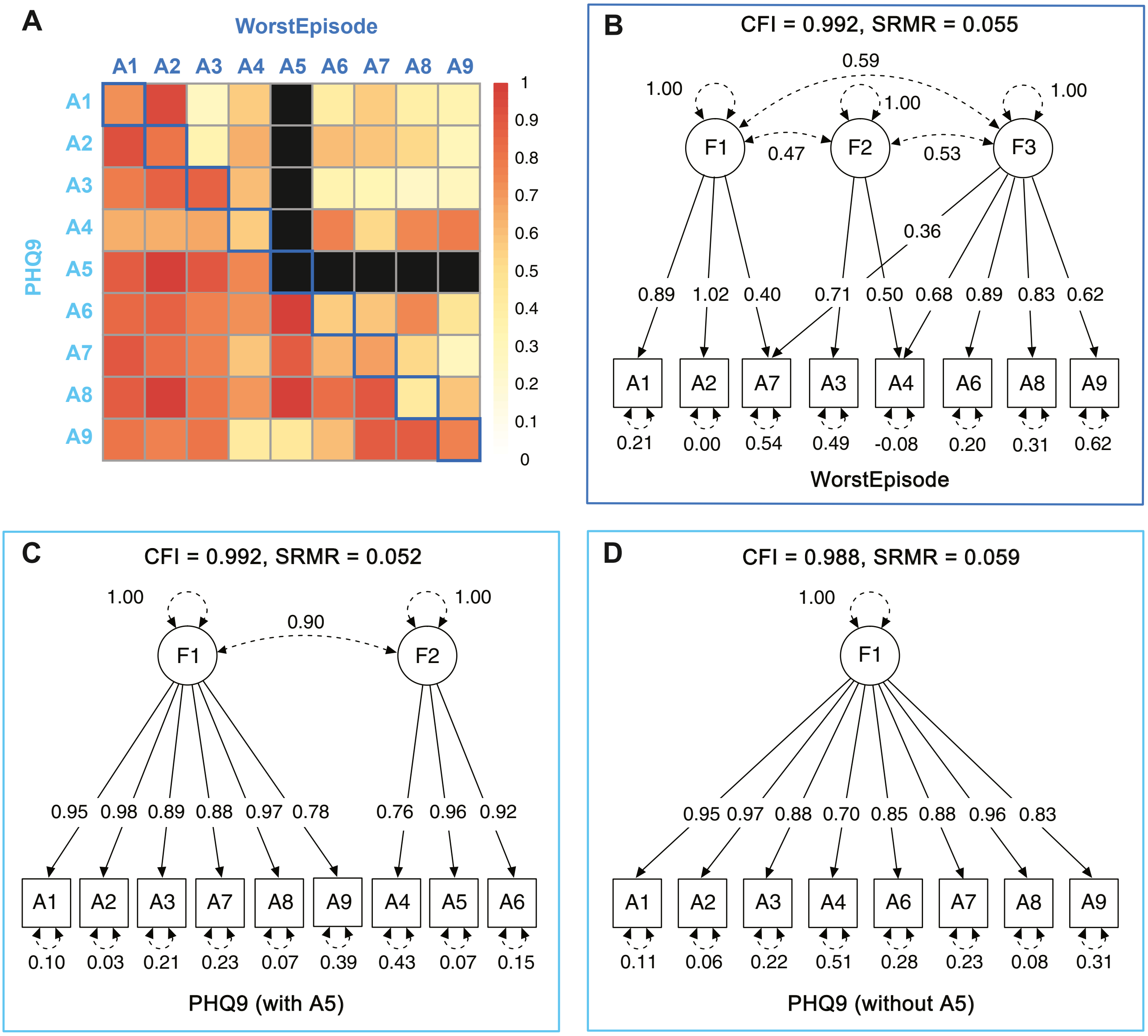


**Figure 6.** Genetic correlation ($r_g$) estimates and factor analysis of Patient Health Questionnaire-9 (PHQ-9) and worst-episode symptoms of major depressive disorder. **(A)** $r_g$ estimates between pairs of symptoms. Blue squares on the diagonal line show $r_g$ estimates between pairs of corresponding worst-episode and PHQ-9 symptoms; squares in the upper triangle show $r_g$ estimates between pairs of worst-episode symptoms; squares in the lower triangle show $r_g$ estimates between pairs of PHQ-9 symptoms. Black squares indicate no data for worst-episode symptom A5. **(B)** Confirmatory factor analysis factor loadings on worst-episode symptoms. **(C)** Confirmatory factor analysis factor loadings on PHQ-9 symptoms, including symptom A5. **(D)** Confirmatory factor analysis factor loadings on PHQ-9 symptoms, excluding symptom A5. Models with the best fits are shown. CFI, comparative fit index; SRMR, standardized root-mean-squared residual.

symptoms lay between 0.41 and 0.89 (Figure S9; Supplemental Methods in Supplement 1).

Finally, we found that the high $r_g$s between PHQ-9 symptoms made them amenable to almost any factor structure. First, we found that a previously proposed 2-factor structure (10,44) fit PHQ-9 symptoms (Akaike information criterion = 91.1, comparative fit index = 0.99, standardized root-mean-squared residual = 0.058) better than worst-episode symptoms (Akaike information criterion = 167.5, comparative fit index = 0.96, standardized root-mean-squared residual = 0.12) (Supplemental Methods in Supplement 1; Table S16 in Supplement 2; Figure S9 in Supplement 1). To verify whether this was specific to the factor structure previously proposed (10,44), we tested all possible unique 2-factor structures for their fits to PHQ-9 symptoms and worst-episode symptoms using CFA, with the only restriction being that the same factor loaded onto symptoms A1 and A2 in each case (Supplemental Methods in Supplement 1). We found that the model fit metrics showed much lower variation for PHQ-9 than for worst-episode symptoms (Figure S9 in Supplement 1).

Therefore, high $r_g$s between the PHQ-9 symptoms made factor structures interchangeable and selection of best-fit models less meaningful.

## DISCUSSION

In this study, we examined whether symptom-level data assessed in the general population by the PHQ-9 captures the same biology as worst-episode MDD symptoms assessed through the CIDI-SF. We found that while they have similar $h^2_{\mathrm{SNP}}$, they have distinct genetic components, and this difference can only be partially accounted for by the skip structure of the CIDI-SF or the severity threshold for symptom endorsement on the PHQ-9. Furthermore, we found that PHQ-9 symptoms were more genetically correlated with each other than worst-episode symptoms were, and factor analysis on their genetic covariance matrices did not identify the same underlying symptom dimensions for MDD. The 2 sets of symptoms are not interchangeable in genetic analyses; they lead to different findings with different biological meanings.

Some of the differences between the 2 sets of symptoms are due to the implementation of the skip structure in the CIDI-SF when assessing worst-episode symptoms. We found that noncardinal worst-episode symptoms were able to capture only those genetic components of MDD that are not shared with the cardinal symptoms. These index different liabilities within individuals enriched for MDD, having the highest PRS pleiotropy for childhood trauma, including violence in the family and having no access to a doctor, as well as items on the GAD7 questionnaire, pointing to them as potential axes of genetic heterogeneity among people with likely MDD. Once the skip structure had been applied to noncardinal PHQ-9 symptoms, most of their differences from the corresponding worst-episode symptoms were gone. The remaining differences may be due to recall differences between current symptoms and symptoms that occurred during a potentially distant MDD episode.

Worst-episode symptoms must, by definition, occur during MDD episodes. Most PHQ-9 symptoms will not, and this can explain most of our results; PHQ-9 symptoms show more genetic sharing with more stable traits like neuroticism, insomnia, and measures of subjective well-being. In other words, PHQ-9 symptoms index general dysphoria more than episodic MDD. Our findings also complement previous findings that self-ratings with the PHQ-9 likely lead to the inclusion of longstanding conditions as well as conditions that are due to external causes unrelated to MDD, which can inflate MDD prevalence (30,31). This may be exacerbated by the healthy volunteer effect in the general population that answers the MHQ in the UK Biobank (2,45). This does not discredit the PHQ-9 as a sensitive screening instrument for current MDD, especially for ruling out those without MDD, or as a measurement of depression severity as it is intended to be (3). However, it may not be suitable for identifying symptom dimensions in patients with MDD. As has been argued previously "not all instruments are appropriate for all purposes" (46,47).

Our results should be interpreted in the context of the following limitations. First, our sample sizes are small, leading to low statistical power in GWAS and MR analyses. This can be improved with increased data collection at the symptom level using methods that go beyond diagnostic questionnaires (47,48). Second, genetic associations identified for PHQ-9 or worst-episode symptoms may be due to collider bias from the ascertainment of individuals participating in the MHQ; participation in the MHQ has positive $r_g$s with higher educational attainment and better health and negative $r_g$s with psychological distress and schizophrenia (45). Third, no clinician ratings are available in the UK Biobank to be compared to PHQ-9 and CIDI-SF ratings, and therefore we do not have insights into biases that are inherent in self-rated symptoms. In particular, self-rated symptoms may suffer from greater recall bias. Both of these limitations may be improved in truly representative, population-based, clinician-assessed cohorts such as electronic health records or national registries, especially when clinicians' notes are available to assess symptom-level disease characteristics. Finally, factors identified through genomic SEM should not necessarily be seen as real or entity-like (10) because they may be subject to statistical underdetermination (49,50). Instead, they reflect genetic sharing among PHQ-9 and worst-episode symptoms. Factor structures for worst-episode symptoms are not only more concordant with previous findings; they are likely more meaningful than those derived from PHQ-9 symptoms, where any factor structure fits equally well.

### Conclusions

In summary, we found that symptoms that were assessed using these 2 instruments captured different underlying biology and were not interchangeable in genetic analysis. PHQ-9 symptoms index general dysphoria more than episodic MDD, and worst-episode symptoms are more suitable for investigations into symptom dimensions of MDD.

## ACKNOWLEDGMENTS AND DISCLOSURES

LH, ST, VZ, and NC are supported by the Technical University of Munich Global Incentive Fund. The iPSYCH team is supported by Lundbeck Foundation (Grants Nos. R102-A9118, R155-2014-1724, and R248-2017-2003); the National Institute of Mental Health (Grant No. 1R01MH124851-01); and the Universities and University Hospitals of Aarhus and Copenhagen. The Danish National Biobank resource is supported by the Novo Nordisk Foundation. High-performance computer capacity for handling and statistical analysis of iPSYCH data on the GenomeDK high-performance computing facility is provided by the Center for Genomics and Personalised Medicine and the Centre for Integrative Sequencing, iSEQ, Aarhus University, Denmark. AJS is supported by Lundbeckfonden (Fellowship No. R335-2019-2318). VA is supported by Lundbeck Foundation postdoctoral (Grant No. R380-2021-1465). ST is supported by a Medical Research Council UK Ph.D. Studentship.

We thank Jonathan Flint, Noah Zaitlen, Joel Mefford, Andrew Dahl, Richard Border, and Eiko Fried for constructive discussions and feedback on the paper.

A previous version of this article was published as a preprint on medRxiv: https://www.medrxiv.org/content/10.1101/2023.02.27.23286527v1.

LH and NC wrote the paper. VZ, KK, and NC designed the study. ST supported the Mendelian randomization analyses. MK, VA, TW, and AJS supported the iPSYCH analyses. LH and NC performed all analyses. All authors reviewed the paper.

This research was conducted under ethical approval from the UK Biobank Resource under application No. 28709. The use of iPSYCH data follows the standards of the Danish Scientific Ethics Committee, the Danish Health Data Authority, the Danish Data Protection Agency, and the Danish Neonatal Screening Biobank Steering Committee. Data access was via secure portals in accordance with Danish data protection guidelines set by the Danish Data Protection Agency, the Danish Health Data Authority, and Statistics Denmark.

UK Biobank genotype and phenotype data used in this study are from the full release of the UK Biobank Resource obtained under application No. 28709. We used publicly available summary statistics from PGC29 (https://www.med.unc.edu/pgc/results-and-downloads). The individual-level data from the iPSYCH cohort are not publicly available due to institutional restrictions on data sharing and privacy concerns. Summary statistics for all PHQ-9 and worst-episode symptoms presented in this paper are available on https://doi.org/10.6084/m9.figshare.22041212. Publicly available tools that are used in data analyses are described wherever relevant in the Methods and Materials and Key Resource Table. Custom code is available at https://github.com/caina89/MDDSymptoms.

The authors report no biomedical financial interests or potential conflicts of interest.

## ARTICLE INFORMATION

From the Helmholtz Pioneer Campus, Helmholtz Zentrum München, Neuherberg, Germany (LH, JR, NC); Computational Health Centre, Helmholtz Zentrum München, Neuherberg, Germany (LH, JR, NC); School of Medicine, Technical University of Munich, Munich, Germany (LH, JR, NC); School of Public Health, Imperial College London, London, United Kingdom (ST, VZ); MRC Centre for Environment and Health, Imperial College London, London, United Kingdom (ST, VZ); Institute of Biological Psychiatry, Mental Health Center, Sct Hans, Copenhagen University Hospital, Mental Health Services CPH, Copenhagen, Denmark (VA, MDK, AJS, TW); Neurogenomics Division, The Translational Genomics Research Institute, Phoenix, Arizona (AJS); Section for Geogenetics, GLOBE Institute, Faculty of Health and Medical Sciences, Copenhagen University, Copenhagen, Denmark (AJS); Lundbeck Foundation GeoGenetics Centre, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark (TW); Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark (TW); and Virginia Institute for Psychiatric and Behavioral Genetics and Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia (KK).

Address correspondence to Na Cai, Ph.D., at na.cai@helmholtz-munich.de.

Received Mar 10, 2023; revised Nov 6, 2023; accepted Nov 26, 2023.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.biopsych.2023.11.021.

## REFERENCES

1. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, *et al.* (2018): The UK biobank resource with deep phenotyping and genomic data. Nature 562:203–209.
2. Davis KAS, Coleman JRI, Adams M, Allen N, Breen G, Cullen B, *et al.* (2020): Mental health in UK Biobank – Development, implementation and results from an online questionnaire completed by 157 366 participants: A reanalysis. BJPsych Open 6:e18.
3. Kroenke K, Spitzer RL, Williams JB (2001): The PHQ-9: Validity of a brief depression severity measure. J Gen Intern Med 16:606–613.
4. American Psychiatric Association, DSM-5 Task Force (2013): Diagnostic and statistical manual of mental disorders: DSM-5. Washington, DC: American Psychiatric Publishing, Inc.
5. Kessler RC, Andrews G, Mroczek D, Ustun B, Wittchen H-U (1998): The World Health Organization Composite International Diagnostic Interview short-form (CIDI-SF). Int J Methods Psychiatr Res 7:171–185.
6. Levinson D, Potash J, Mostafavi S, Battle A, Zhu X, Weissman M (2017): Brief assessment of major depression for genetic studies: Validation of Cidi-Sf screening with Scid interviews. Eur Neuropsychopharmacol 27:S448.
7. Thorp JG, Marees AT, Ong JS, An J, MacGregor S, Derks EM (2020): Genetic heterogeneity in self-reported depressive symptoms identified through genetic analyses of the PHQ-9. Psychol Med 50: 2385–2396.
8. Grotzinger AD, Rhemtulla M, de Vlaming R, Ritchie SJ, Mallard TT, Hill WD, *et al.* (2019): Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nat Hum Behav 3:513–525.
9. Thorp JG, Campos AI, Grotzinger AD, Gerring ZF, An J, Ong JS, *et al.* (2021): Symptom-level modelling unravels the shared genetic architecture of anxiety and depression. Nat Hum Behav 5:1432–1442.
10. van Loo HM, Aggen SH, Kendler KS (2022): The structure of the symptoms of major depression: Factor analysis of a lifetime worst episode of depressive symptoms in a large general population sample. J Affect Disord 307:115–124.
11. Li Y, Aggen S, Shi S, Gao J, Li Y, Tao M, *et al.* (2014): The structure of the symptoms of major depression: Exploratory and confirmatory factor analysis in depressed Han Chinese women. Psychol Med 44:1391–1401.
12. Kendler KS, Aggen SH, Neale MC (2013): Evidence for multiple genetic factors underlying DSM-IV criteria for major depression. JAMA Psychiatry 70:599–607.
13. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015): Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience 4:7.
14. Abraham G, Qiu Y, Inouye M (2017): FlashPCA2: Principal component analysis of biobank-scale genotype datasets. Bioinformatics 33: 2776–2778.
15. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, *et al.* (2018): Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet 50:668–681.
16. Mors O, Perto GP, Mortensen PB (2011): The Danish psychiatric central research register. Scand J Public Health 39(suppl):54–57.
17. Lynge E, Sandegaard JL, Rebolj M (2011): The Danish national patient register. Scand J Public Health 39(suppl):30–33.
18. Pedersen CB, Bybjerg-Grauholm J, Pedersen MG, Grove J, Agerbo E, Bækvad-Hansen M, *et al.* (2018): The iPSYCH2012 case-cohort sample: New directions for unravelling genetic and environmental architectures of severe mental disorders. Mol Psychiatry 23:6–14.
19. Bybjerg-Grauholm J, Bøcker Pedersen C, Bækvad-Hansen M, Giørtz Pedersen M, Adamsen D, Søholm Hansen C, *et al.* (2020, December 2): The iPSYCH2015 Case-Cohort sample: Updated directions for unravelling genetic and environmental architectures of severe mental disorders. medRxiv. https://doi.org/10.1101/2020.11.30.20237768.
20. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, *et al.* (2015): LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47:291–295.
21. Yavorska OO, Burgess S (2017): MendelianRandomization: An R package for performing Mendelian randomization analyses using summarized data. Int J Epidemiol 46:1734–1739.
22. Burgess S, Foley CN, Zuber V (2018): Inferring causal relationships between risk factors and outcomes from genome-wide association study data. Annu Rev Genomics Hum Genet 19:303–327.
23. Bowden J, Davey Smith G, Haycock PC, Burgess S (2016): Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol 40:304–314.
24. Bowden J, Davey Smith G, Burgess S (2015): Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. Int J Epidemiol 44:512–525.
25. Morrison J, Knoblauch N, Marcus JH, Stephens M, He X (2020): Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nat Genet 52:740–747.
26. Zuber V, Colijn JM, Klaver C, Burgess S (2020): Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. Nat Commun 11:29.
27. Choi SW, O'Reilly P (2019): PRSice 2: Polygenic risk score software (updated) and its application to cross-trait analyses. Eur Neuropsychopharmacol 29:S832.
28. Dahl A, Thompson M, An U, Krebs MD, Appadurai V, Border R, *et al.* (2023): Phenotype integration improves power and preserves

specificity in biobank-based genetic studies of major depressive disorder. Nat Genet 55:2082–2093.

29. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, *et al.* (2015): An atlas of genetic correlations across human diseases and traits. Nat Genet 47:1236–1241.

30. von Glischinski M, von Brachel R, Thiele C, Hirschfeld G (2021): Not sad enough for a depression trial? A systematic review of depression measures and cut points in clinical trial registrations. J Affect Disord 292:36–44.

31. Levis B, Benedetti A, Ioannidis JPA, Sun Y, Negeri Z, He C, *et al.* (2020): Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: Individual participant data meta-analysis. J Clin Epidemiol 122:115–128.e1.

32. Havinga PJ, Boschloo L, Bloemen AJP, Nauta MH, de Vries SO, Penninx BWJH, *et al.* (2017): Doomed for disorder? High incidence of mood and anxiety disorders in offspring of depressed and anxious patients: A prospective cohort study. J Clin Psychiatry 78:e8–e17.

33. Kendler KS, Gardner CO, Gatz M, Pedersen NL (2007): The sources of co-morbidity between major depression and generalized anxiety disorder in a Swedish national twin sample. Psychol Med 37:453–462.

34. Kendler KS, Karkowski LM, Prescott CA (1999): Causal relationship between stressful life events and the onset of major depression. Am J Psychiatry 156:837–841.

35. Kendler KS, Karkowski-Shuman L (1997): Stressful life events and genetic liability to major depression: Genetic control of exposure to the environment? Psychol Med 27:539–547.

36. Peterson RE, Cai N, Dahl AW, Bigdeli TB, Edwards AC, Webb BT, *et al.* (2018): Molecular genetic analysis subdivided by adversity exposure suggests etiologic heterogeneity in major depression. Am J Psychiatry 175:545–554.

37. Eysenck SBG, Eysenck HJ, Barrett P (1985): A revised version of the psychoticism scale. Pers Individ Dif, rev. version 6:21–29.

38. Burgess S, Davey Smith G, Davies NM, Dudbridge F, Gill D, Glymour MM, *et al.* (2019): Guidelines for performing Mendelian randomization investigations: Update for summer 2023. Wellcome Open Res 4:186.

39. Smith GD, Ebrahim S (2003): “Mendelian randomization”: Can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 32:1–22.

40. Cai N, Revez JA, Adams MJ, Andlauer TFM, Breen G, Byrne EM, *et al.* (2020): Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet 52:437–447.

41. Shafer AB (2006): Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. J Clin Psychol 62:123–146.

42. Romera I, Delgado-Cohen H, Perez T, Caballero L, Gilaberte I (2008): Factor analysis of the Zung Self-Rating Depression Scale in a large sample of patients with major depressive disorder in primary care. BMC Psychiatry 8:4.

43. van Loo HM, de Jonge P, Romeijn JW, Kessler RC, Schoevers RA (2012): Data-driven subtypes of major depressive disorder: A systematic review. BMC Med 10:156.

44. Elhai JD, Contractor AA, Tamburrino M, Fine TH, Prescott MR, Shirley E, *et al.* (2012): The factor structure of major depression symptoms: A test of four competing models using the Patient Health Questionnaire-9. Psychiatry Res 199:169–173.

45. Adams MJ, Hill WD, Howard DM, Dashti HS, Davis KAS, Campbell A, *et al.* (2020): Factors associated with sharing e-mail information and mental health survey participation in large population cohorts. Int J Epidemiol 49:410–421.

46. Patalay P, Fried EI (2021): Editorial Perspective: Prescribing measures: Unintended negative consequences of mandating standardized mental health measurement. J Child Psychol Psychiatry 62:1032–1036.

47. Fried EI, Flake JK, Robinaugh DJ (2022): Revisiting the theoretical and methodological foundations of depression measurement. Nat Rev Psychol 1:358–368.

48. Freimer NB, Mohr DC (2019): Integrating behavioural health tracking in human genetics research. Nat Rev Genet 20:129–130.

49. Johnson K (2016): Realism and uncertainty of unobservable common causes in factor analysis. Nous 50:329–355.

50. Romeijn JW, Williamson J (2018): Intervention and identifiability in latent variable modelling. Minds Mach (Dordr) 28:243–264.