Methods

iPSYCH 2015 case-cohort study

The Lundbeck Foundation initiative for Integrative Psychiatric Research (iPSYCH)[1,2] is a case-cohort study of all singleton births between 1981 and 2008 to mothers legally residing in Denmark and who were alive and residing in Denmark on their first birthday (N=1,657,449). The iPSYCH 2015 case-cohort comprises two enrollments from this base population. The iPSYCH 2012 case-cohort enrolled 86,189 individuals (30,000 random population controls; 57,377 psychiatric cases)[1]. The iPSYCH 2015i case-cohort expanded enrollment by an additional 56,233 individuals (19,982 random population controls; 36,741 psychiatric cases)[1,2]. DNA was extracted from dried blood spots stored in the Danish Neonatal Screening Biobank[3] and genotyping was performed on the Infinium PsychChip v1.0 array (2012) or the Global Screening Array v2 (2015i). Psychiatric diagnoses were obtained from the Danish Psychiatric Central Research Register (PCR)[4] and the Danish National Patient Register (DNPR)[5]. Diagnoses in these registers are made by licensed psychiatrists during in- or out- patient specialty care but diagnoses or treatments assigned in primary care are not included. Linkage across population registers, to parents where known, and to the neonatal biobank is possible via unique citizen identifiers of the Danish Civil Registration System[6]. The use of this data follows 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.

Genotyping and quality control

Genotype phasing, imputation, and quality control were performed in parallel in the 2012 and 2015i cohorts according to custom, mirrored protocols. Briefly, phasing and imputation were conducted using BEAGLEv5.1[7,8], both steps including reference haplotypes from the Haplotype Reference Consortium v1.1 (HRC)[9]. Quality control was applied prior to and following imputation to correct for missing data across SNPs and individuals, SNPs showing deviations from Hardy-Weinberg equilibrium in cases or controls, abnormal heterozygosity of SNPs and samples, genotype-phenotype sex discordance, minor allele frequency (MAF), batch artifacts, and imputation quality. Kinship was detected within and across 2012 and 2015i cohorts using KING[10], censoring to ensure no second degree or high relatives remained. Ancestry was examined using the smartpca module of EIGENSOFT[11], and multivariate PCA outliers from the set of iPSYCH individuals with both grandparents and four grandparents born in

Denmark were excluded. In total, 7,649,999 imputed allele dosages were retained for analysis (MAF > 0.01).

iPSYCH 2015 case-cohort genealogies

All recorded relatives of probands in this iPSYCH 2015 case-cohort were obtained from the Danish Civil Registry[6] using mother-father-offspring linkages. From the 141,265[12] probands, we identified 2,066,657 unique relatives, assembling all relationships into a population graph using the *kinship2*[13] and *FamAgg*[12] packages where edges denoted membership in a recorded trio. The relatedness coefficient for each pair was calculated as a weighted sum of unique ancestral paths through the population graph (i.e. not including the same individual, except for the common ancestor). Each path in the sum was weighted by (0.5)*(number of edges in the path)[14]. The Danish Civil Registry does not contain information on zygosity for same-sex twins, but following analysis of the SNP-kinship of children of same-sex twins (Supplementary Figure 3) we assigned same-sex twins a relatedness coefficient of 0.75. Similarly, guided by analysis of siblings with missing paternal records (Supplementary Figure 2), we assigned maternal siblings with missing paternal records a relatedness coefficient of 0.25. 24,773 pairs of relatives from the population genealogy included two probands genotyped on the same genotype array. We used Pearson's correlation of the graph-inferred kinship and SNP-inferred kinship using KING[10] as an estimate of concordance and quality of inferred relationships.

Pearson-Aitken Family Genetic Risk Scores (PA-FGRS)

PA-FGRS estimates a liability for disease carried by a proband from the observed disease status in a pedigree and under the assumption of a liability threshold model for the disease[15]. The method first estimates an initial liability for each relative and then uses the Pearson-Aitken selection formula to sequentially update the expected liability in the proband conditional on each relative.[15,16]

We begin by assuming a disease, $D_i = 1$, arises when an individual, i, carries a latent liability, L_i, that surpasses some threshold, t. Liability, L_i, can arise from additive effects (β_j) of genetic factors (X_{ij}), or environmental deviations (e_i) and genetic contributions follow classic polygenic theory.[15,16] We can write a generative model:

$$L_i = \sum_j \beta_j X_{ij} + e_i$$

$$D_i = \begin{cases} 1\,, & L_i \geq t \\ 0\,, & L_i < t \end{cases}, \quad t = \Phi^{-1}\left(1 - K_{pop}\right)$$

Where the threshold, t, is the standard normal quantile that corresponds to a cumulative probability of k_{pop}, the lifetime prevalence of the disorder. Further we assume that the vector consisting of the genetic liability of the proband and the total liability of n genetic relatives $\left[G_p, L_1, \ldots, L_n\right]^T \sim MVN([0, \ldots, 0]^T, \Sigma)$ with covariance matrix:

$$\Sigma = \begin{bmatrix} h_l^2 & h_l^2 r_{p,1} & \cdots & h_l^2 r_{p,n} \\ h_l^2 r_{p,1} & 1 & \cdots & h_l^2 r_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ h_l^2 r_{p,n_{rel}} & h_l^2 r_{1,n} & \cdots & 1 \end{bmatrix}$$

Under this model the expected value of L_i, conditional on the true value for D_i is according to truncated normal distribution theory[17]:

$$\mathrm{E}\left(L_i | D_i\right) = \begin{cases} \dfrac{-\varphi(t)}{k_{pop}}, & D_i = 0 \\ \dfrac{\varphi(t)}{1 - k_{pop}} & D_i = 1 \end{cases}$$

A critical assumption of this model is that each individual is fully observed, meaning there is an equivalence between their diagnostic and disorder status. This assumption rarely holds in practice, but the variable follow-up of relatives by the Danish register system makes it extremely tenuous. We instead propose a model where the disease status Y_i in those who surpass the threshold is a stochastic process with a probability corresponding to the ratio between the age-specific prevalence (K_i) and the life-time prevalence (K_{pop})

$$Y_i = \begin{cases} Bernoulli\left(\dfrac{K_i}{K_{Pop}}\right), & D_i = 1 \\ 0 \qquad\qquad\quad , & D_i = 0 \end{cases}$$

To get the expected liabilities under this model we use a mixture of an upper and a lower truncated Gaussian both with mean and variance corresponding to their conditional expectations, and with the mixture proportion (π_n), corresponding to the conditional probability of being a case. Let $\psi(\mu, \sigma^2, a, b)$ denote a truncated gaussian with mean μ, variance σ^2, lower truncation at a and upper truncation at b. Then the distribution of L_n conditional observations 1 to n is:

$$L_n|Y_1,\ldots,Y_n,K_1,\ \ldots,\ K_n,K_{pop},\Sigma\ \sim \pi_n\psi\left(\mu_n,\ \Omega_{x,x}a = t, b = \infty\right) + \left(1 - \pi_n\right)\psi\left(\mu_n,\ \Omega_{x,x}, a = -\infty, b = t\right)$$

with $\mu_n = 0$ if n=1 and $\mu_n = \mathrm{E}(L_n|Y_1,\ldots,Y_{n-1},K_1,\ \ldots\ K_{n-1},K_{pop},\Sigma)$ otherwise, while $\Omega_{x,x} = 1$ if n=1 and $\Omega_{x,x} = \mathrm{Var}(L_n|Y_1,\ldots,Y_{n-1},K_1,\ \ldots\ K_{n-1},K_{pop},\Sigma)$ otherwise (see below), and $\pi_n = 1$ if $Y_n = 1$ and $\pi_n = \mathrm{P}(D_n = 1|Y_1,\ldots,Y_n,K_1,\ \ldots\ K_n,K_{pop},\Sigma)$ otherwise. This we approximate as:

$$\mathrm{P}(D_n = 1|Y_1,\ldots,Y_{n-1},Y_n = 0,K_1,\ \ldots\ ,K_n,K_{pop},\Sigma) \approx 1 - \frac{\Phi(\frac{T-\mu_n}{\sqrt{\Omega_{n,n}}})}{\Phi(\frac{T-\mu_n}{\sqrt{\Omega_{n,n}}}) + \frac{K_{pop}-K_n}{K_{pop}}(1-\Phi(\frac{T-\mu_n}{\sqrt{\Omega_{n,n}}}))}$$

Following adaptations[18,19] of the Pearson-Aitken selection formula[20] the conditional mean and variance of expected liability for a proband is estimated given their pedigree, initial liabilities, and population parameters[19]. Let $\mu^*_n - \mu_n$ be the effect conditioning Y_n and K_n has on μ_n then the vector of conditional mean liabilities *(μ*)* is::

$$\mu^* = \mu + \Omega_{,x}\Omega_{x,x}^{-1}(\mu^*_x - \mu_x) \qquad \text{Eq. 1}$$

Where Ω is the covariance matrix of the liabilities. Similarly, if conditioning change $\Omega_{x,x}$ to $\Omega^{*}_{x,x}$, the conditional covariance matrix of liabilities, Ω^{*}, is estimated as:

$$\Omega^{*} = \Omega - \Omega_{,x}\left(\Omega_{x,x}^{-1} - \Omega_{x,x}^{-1}\Omega^{*}_{x,x}\Omega_{x,x}^{-1}\right)\Omega_{x,}$$

Previous work has found this to be an efficient estimator of genetic liabilities of binary traits given family history.[18,19,21] In practice, we start by setting the liability vector to a zero vector, we then iteratively condition on the observed disease status of each relative using the expected mean and a variance of a mixture of truncated gaussians in combination with the Pearson-Aitken selection formula to obtain the expected genetic liability of the index individual.

Our PA-FGRS is available as R code: https://github.com/MortenKrebs/PA-FGRS.

Simulations

We simulated pedigrees for 500 probands with different family histories including varying pedigree structure and censoring. The heritability was set to 0.50 and prevalence to 0.4. We assessed the correlation between the estimated liabilities obtained from five different liability estimation methods. Next, we repeated the simulations 4000 times with varying prevalence and 5000 times with varying heritability.

To assess the impact of shared environment (c^2), we simulated varying levels of shared environment between parents, off-spring and siblings. We estimated the correlation of FGRS and the true genetic and environmental liability. For $FGRS_{Kendler}$[22] we included both an c^2-adjusted and an unadjusted version. If shared environment effects are expected we propose a modification to PA-FGRS, $PA\text{-}FGRS_{noFDR}$ that omits parents, siblings and children when estimating liability. We used simulations to compare this approach to FGRS and the full pedigree PA-FGRS, computing the correlation between true and estimated liability.

Psychiatric phenotypes

Our primary outcome, MDD, was defined as having a registration with a depressive episode (F32) or recurrent depression (F33) before Jan 1st 2017, according, the Danish Psychiatric Central Research Register (PCR)[4]. Diagnostic codes used for the construction of PA-FGRS scores are found in

Supplementary Table S1. For relatives diagnosed between 1968 and 1994 records are limited to in-patient contacts and ICD-8 codes.

Population parameters used for computing PA-FGRS in iPSYCH

The sex-specific lifetime prevalence of each disorder (Supplementary Table S1) was obtained from published estimates based on Danish registers[23] . Heritability parameters were estimated chosen from literature (Supplementary Table S1). Sex and birth-year-specific cumulative incidence curves were computed based on a sample consisting of all members of the iPSYCH-2015 random sample and all their available relatives (N=979,582; Supplementary Figure S14).

Polygenic Scores

PGS for MDD, SCZ and BP were computed based on published, external summary statistics (Supplementary Table S2) that had no overlap with iPSYCH, while PGS for ASD and ADHD were based on GWAS run in other half of the unrelated subset of iPSYCH (iPSYCH2012 for iPSYCH2015i and vice versa, Supplementary Figure S9) . We used SBayesR [24] to compute SNP-weights for SNPs in the intersection of each GWAS and iPSYCH. Palindromic SNPs (A/T, C/G), those not mapping uniquely to hg19 positions, and without a unique rsID in dbSNP v151 were excluded.

Classification analysis

In the European subset of the iPSYCH-2015-MDD case-cohort (Supplementary Figure S9), we used logistic regressions with MDD as an outcome and first using either PA-FGRS$_{MDD}$, PGS$_{MDD}$ or both of these as predictors, and afterwards using either five PA-FGRSs, five PGSs, or all ten of these, the five scores corresponding to the selected categories of mental disorders (Supplementary Table S1-S2). PA-FGRS for these comparisons were estimated blind to the proband status. The classification accuracy was assessed using the area under the receiver operating characteristic curve (AUC) estimated using the *pROC* package[25].

Comparing Polygenic profiles

Among the individuals diagnosed with MDD, putative subgroup-defining features were obtained from the PCR[4] and the Danish Civil Registry[6]: a diagnosis of BPD (ICD10: F30-F31), comorbid anxiety (F40.0-40.2, F41.0-41.1, or F42), sex (as registered at birth), recurrence (ICD10: F32 or F33), severity (ICD10: F32/33.0, F32/33.1, F32/33.2, or F32/33.3), age at first recorded diagnosis, and mode of

treatment (inpatient, casualty-ward or outpatient). We computed a composite estimate of genetic liability for each of the five mental disorders as a weighted sum of the PGS and PA-FGRS with weights corresponding to the betas from a logistic regression of their natural outcome in a calibration sample (Supplementary Figure S9). For each outcome, multiple multinomial logistic regression was fitted to sequentially estimate the effects of each the composite genetic risk estimates with age, sex and 10 genetic PCs as covariates using the R package *nnet*[26]. We report a normalized partial effect size for each PGS and FGRS, (β_{MLR}/β_{LR}) which is the ratio of the effect of the PA-FGRS on MDD outcomes (β_{MLR}) over its effect on the natural outcomes (β_{LR}; e.g., ASD for FGRS for ASD) where the β_{LR} were estimated in outcome-specific case cohort samples (e.g. ASD case cohort, Supplementary Figure S9). This was done to enable intuitive effect size comparisons of each predictor on the various outcomes. These analyses were conducted separately for iPSYCH-2012 and iPSYCH-2015 samples and meta-analyzed. Subgroup-level effect estimates were meta-analyzed using inverse variance weighting, while heterogeneity test p–values were combined using Fisher's method. In total we report 35 p-values declaring 0.05/35 = 0.0014 strictly significant.

Genome-Wide Association Studies (GWAS)

GWAS were performed within two proband groups, the iPSYCH2012 MDD case-cohort and the iPSYCH2015i MDD case-cohort, on imputed allelic dosage data using plink2[27]. For binary MDD diagnosis, logistic regression was applied, for continuous valued PA-FGRS, we used linear regression, both including sex and age and 10 principal components of genetic ancestry as covariates. Inverse-variance weighted meta-analysis of the two constituent samples was performed using METAL[28]. SNPs with association p-values less than $5x10^{-8}$ were declared significant, while variants with a false discovery rate of 0.05 were considered suggestive. Loci were considered independent if >1Mb apart. Observed-scale SNP-heritability ($h^2_{SNP,obs}$) and genetic correlations to nine published GWAS (Supplementary Table S3) were estimated using LDscore regression[29,30]. Difference in $h^2_{SNP,obs}$ was computed as $h^2_{PA-FGRS} - h^2_{case/ctrl}$, with std.err. $\approx \sqrt{s.e.(h^2_{PA-FGRS})^2 + s.e.(h^2_{case/ctrl})^2}$. Genome-wide significant index SNPs were defined from a large external GWAS of MDD, modified to exclude 23andMe and iPSYCH, by clumping overlapping SNP lists. A paired t-test of the squared test statistic was used to assess significance of improvement. Polygenic scores for within iPSYCH classification were computed using SNPs with MAF>0.01 and INFO>0.8, clumped and thresholded with Plink 1.90b6.27[27], using parameters --clump-kb 625 --clump-p1

0.1 --clump-p2 0.1 --clump-r2 0.8. Improvements in predictions were assessed using the difference in AUC test in the pROC package.

Online References

1. Pedersen, C. B. *et al.* The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
2. Bybjerg-Grauholm, J. *et al.* The iPSYCH2015 Case-Cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders. *bioRxiv* (2020) doi:10.1101/2020.11.30.20237768.
3. Nørgaard-Pedersen, B. & Hougaard, D. M. Storage policies and use of the Danish Newborn Screening Biobank. *J. Inherit. Metab. Dis.* **30**, 530–536 (2007).
4. Mors, O., Perto, G. P. & Mortensen, P. B. The Danish Psychiatric Central Research Register. *Scand. J. Public Health* **39**, 54–57 (2011).
5. Lynge, E., Sandegaard, J. L. & Rebolj, M. The Danish National Patient Register. *Scand. J. Public Health* **39**, 30–33 (2011).
6. Pedersen, C. B. The Danish Civil Registration System. *Scand. J. Public Health* **39**, 22–25 (2011).
7. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
8. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
9. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
10. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
11. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

12. Rainer, J. *et al.* FamAgg: an R package to evaluate familial aggregation of traits in large pedigrees. *Bioinformatics* **32**, 1583–1585 (2016).
13. Sinnwell, J. P., Therneau, T. M. & Schaid, D. J. The kinship2 R package for pedigree data. *Hum. Hered.* **78**, 91–93 (2014).
14. Wright, S. Coefficients of Inbreeding and Relationship. *Am. Nat.* **56**, 330–338 (1922).
15. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).
16. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
17. Johnson, N. L., Kotz, S. & Balakrishnan, N. *Continuous Univariate Distributions, Volume 2*. (John Wiley & Sons, 1995).
18. So, H.-C., Kwan, J. S. H., Cherny, S. S. & Sham, P. C. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am. J. Hum. Genet.* **88**, 548–565 (2011).
19. Mendell, N. R. & Elston, R. C. Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. *Biometrics* **30**, 41–57 (1974).
20. Aitken, A. C. Note on selection from a multivariate normal population. *Proc. Edinb. Math. Soc.* (1935).
21. Hujoel, M. L. A., Gazal, S., Loh, P.-R., Patterson, N. & Price, A. L. Liability threshold modeling of case–control status and family history of disease increases association power. *Nat. Genet.* **52**, 541–547 (2020).
22. Kendler, K. S., Ohlsson, H., Sundquist, J. & Sundquist, K. Family Genetic Risk Scores and the Genetic Architecture of Major Affective and Psychotic Disorders in a Swedish National Sample. *JAMA Psychiatry* **78**, 735–743 (2021).

23. Pedersen, C. B. *et al.* A comprehensive nationwide study of the incidence rate and lifetime risk for treated mental disorders. *JAMA Psychiatry* **71**, 573–581 (2014).
24. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).
25. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
26. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer Science & Business Media, 2003).
27. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
28. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
29. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
30. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).