Archival Report

Biological Psychiatry

A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts

Guiyan Ni, Jian Zeng, Joana A. Revez, Ying Wang, Zhili Zheng, Tian Ge, Restuadi Restuadi, Jacqueline Kiewa, Dale R. Nyholt, Jonathan R.I. Coleman, Jordan W. Smoller, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Jian Yang, Peter M. Visscher, and Naomi R. Wray

ABSTRACT

BACKGROUND: Polygenic scores (PGSs), which assess the genetic risk of individuals for a disease, are calculated as a weighted count of risk alleles identified in genome-wide association studies. PGS methods differ in which DNA variants are included and the weights assigned to them; some require an independent tuning sample to help inform these choices. PGSs are evaluated in independent target cohorts with known disease status. Variability between target cohorts is observed in applications to real data sets, which could reflect a number of factors, e.g., phenotype definition or technical factors.

METHODS: The Psychiatric Genomics Consortium Working Groups for schizophrenia and major depressive disorder bring together many independently collected case-control cohorts. We used these resources (31,328 schizophrenia cases, 41,191 controls; 248,750 major depressive disorder cases, 563,184 controls) in repeated application of leave-one-cohort-out meta-analyses, each used to calculate and evaluate PGS in the left-out (target) cohort. Ten PGS methods (the baseline PC+T method and 9 methods that model genetic architecture more formally: SBLUP, LDpred2-Inf, LDpred-funct, LDpred2, Lassosum, PRS-CS, PRS-CS-auto, SBayesR, MegaPRS) were compared.

RESULTS: Compared with PC+T, the other 9 methods gave higher prediction statistics, MegaPRS, LDPred2, and SBayesR significantly so, explaining up to 9.2% variance in liability for schizophrenia across 30 target cohorts, an increase of 44%. For major depressive disorder across 26 target cohorts, these statistics were 3.5% and 59%, respectively.

CONCLUSIONS: Although the methods that more formally model genetic architecture have similar performance, MegaPRS, LDpred2, and SBayesR rank highest in most comparisons and are recommended in applications to psychiatric disorders.

https://doi.org/10.1016/j.biopsych.2021.04.018

Polygenic scores (PGSs), which assess the genetic risk of individuals for a disease (1,2), are calculated as a weighted count of genetic risk alleles in the genome of an individual, with the risk alleles and their weights derived from the results of genome-wide association studies (GWASs) (3). PGSs can be calculated for any trait or disease with sufficiently powered GWASs (discovery samples), and accuracy of PGSs applied in independent GWAS target samples will increase as discovery sample size increases. As genetic factors capture only the genetic contribution to risk and as PGSs capture only part of the genetic risk, PGSs cannot be diagnostically accurate risk predictors [see review (4)]. Nonetheless, for many common complex genetic disorders, such as cancers (5,6) and heart disease (7,8), there is increasing interest in evaluating PGSs for early disease detection, prevention, and intervention (9–11).

There are now many methods to calculate PGSs, and the methods differ in terms of two key criteria: which DNA variants to include and what weights to allocate to them. In this article, for simplicity, we assume that the DNA variants are single nucleotide polymorphisms (SNPs), but other DNA variants tested for association with a trait can be used. While stringent thresholds are set to declare significance for association of individual SNPs in GWASs, PGSs are robust to inclusion of some false positives. Hence, the maximum prediction from PGSs tested in target samples may include nominally associated SNPs. The optimal method to decide which SNPs to select and what weights to allocate to them may differ among traits depending on the sample size of the discovery GWAS and on the genetic architecture of the trait (the number, frequencies, and effect sizes of causal variants), particularly given

ISSN: 0006-3223

the linkage disequilibrium (LD) correlation structure between SNPs. Often, when new PGS methods are introduced, comparisons are made between a limited set of methods using simulated data, together with application to some real data examples. However, it can be difficult to compare across the new methods, particularly because in real data there can be variability in PGS evaluation statistics between target cohorts that is not encountered in idealized simulations. The reasons for this variability are usually unknown and not simple to identify (12) but could reflect a number of factors, such as phenotype definition, ascertainment strategies of cases and controls, cohort-specific ancestry within the broad classification of ancestry defined by the GWAS discovery samples (e.g., European), or technical artifacts in genotype generation.

We compared 10 PGS methods [PC+T (3,13), SBLUP (14), LDpred2-Inf (15), LDpred2 (15), LDpred-funct (16), Lassosum (17), PRS-CS (18), PRS-CS-auto (18), SBayesR (19), and MegaPRS (20)] (Table 1). Some of these methods (PC+T, LDpred2, MegaPRS, Lassosum, and PRS-CS) require a tuning sample, a GWAS cohort with known trait status that is independent of both discovery and target samples, used to select parameters needed to generate the PGSs in the target sample. Whereas only GWAS summary statistics are needed for discovery samples, individual-level genotype data are needed for tuning and target samples. Information about the LD structure is supplied by a reference data set of genome-wide genotypes that can be from independently collected GWAS data of matched ancestry.

Briefly, PC+T (*p* value–based clumping and thresholding, also known as P+T or C+T) uses the GWAS effect size estimates as SNP weights and includes independent SNPs (defined by an LD r^2 filter for a given chromosomal window distance) with association *p* values lower than a threshold (chosen after application in a tuning sample). PC+T is the most commonly used and basic method and so is the benchmark method here. Other methods assume either that all SNPs have an effect size drawn from a normal distribution (SBLUP and LDpred2-Inf) or that SNP effects are drawn from mixtures of distributions with the key parameters defining these architectures estimated through Bayesian frameworks (LDpred2, PRS-CS, SBayesR). The methods LDpred-funct and MegaPRS include functional annotation to SNPs to up- or downweight their contributions to the PGSs, which could improve prediction accuracy if this functional information helps to better separate true- and false-positive associations (21). The MegaPRS software implements a suite of methods (Table 1) and selects the method, together with its parameter estimates, that maximizes prediction in the tuning cohort. MegaPRS uses the BLD-LDAK model (22), in which the variance explained by each SNP depends on its allele frequency, LD, and functional annotations. Notably, some methods (SBayesR, PRS-CS-auto, and LDpred2-auto) do not require a tuning cohort, so that the SNPs selected and their weights reflect only the properties of the discovery sample. As LDpred2-auto is shown to perform similarly to LDpred2, we did not include it in comparisons made here. We applied these methods to data from the Psychiatric Genomics Consortium (PGC) Working Groups for schizophrenia (SCZ) (23) and major depressive disorder (MDD) (Tables S1 and S2 in Supplement 2) (12,24,25). We selected SCZ and MDD to study as they have the largest GWAS samples for psychiatric disorders to date but are diverse in lifetime risk and are representative of all psychiatric disorders, which have been shown to be highly polygenic (26). The PGC provides a useful resource for undertaking this study because it brings together many independently collected cohorts for GWAS meta-analysis. This allows the application of repeated leave-one-cohort-out GWAS analyses, generating robust conclusions from evaluation of PGSs applied across multiple left-out target cohorts.

METHODS AND MATERIALS

Data

All samples were of European ancestry; see full details in Supplement 1 and Tables S1 and S2 in Supplement 2. Briefly, GWAS summary statistics were available from the PGC SCZ Working Group for 37 European ancestry cohorts (23) (31,328 SCZ cases and 41,191 controls), of which 34 had individual-level data available. PGSs were calculated in each of the 30 cohorts (target samples) using the GWAS discovery sample based on a meta-analysis of 37 − 2 = 35 cohorts (23) (the target sample was excluded from the discovery sample as well as a sample selected to be a tuning sample). Analyses were repeated using 4 different tuning samples, 2 large (swe6: 2313; gras: 2318) and 2 small (lie2: 406; msaf: 466). Similarly, GWAS MDD summary statistics were available from 248,750 cases and 563,184 controls (24), which included data from the 26 cohorts from the PGC MDD Working Group with individual-level data (15,805 cases and 23,340 controls). We left one cohort out of those 26 cohorts in turn as the target sample and then used a meta-analysis of remaining data as discovery samples. A cohort (24) that was not included in the discovery GWAS was used as the tuning sample (N = 1679).

Baseline SNP Selection

For baseline analyses, only SNPs with minor allele frequency (MAF) > 0.1 and imputation quality INFO score > 0.9 (converted to best-guess genotype values of 0, 1, or 2) were selected. Sensitivity analyses relaxed the MAF threshold to MAF > 0.05 or 0.01 and INFO score threshold to 0.3. All methods were conducted using HapMap3 SNPs (27) except PC+T, which was conducted based on all imputed SNPs (8 million in SCZ and 13 million in MDD).

Prediction Methods

We defined a PGS of an individual, j, as a weighted sum of SNP allele counts: $\sum_{i=1}^{m} \hat{b}_i x_{ij}$, where m is the number of SNPs included in the predictor, $\hat{b}_i$ is the per allele weight for the SNP, x_{ij} is a count of the number (0, 1, or 2) of trait-associated alleles of SNP i in individual j. We compared 10 risk prediction methods, described in Prediction Methods in Supplement 1 and summarized in Table 1. The methods differ in terms of the SNPs selected for inclusion in the predictor and the $\hat{b}_i$ values assigned to the SNPs. All methods use the GWAS summary statistics as the starting point, but each makes choices differently for which SNPs to include and for the $\hat{b}_i$ values to assign. Some methods use a tuning cohort; parameter estimates that maximize prediction in that tuning cohort

Table 1. Summary of Methods Used to Generate Polygenic Scores

Method	Distribution of SNP Effects (β)	Tuning Sample	Predefined Parameters	Parameters Estimated in Tuning Sample
PC+T	None	Yes	–	*p*-value threshold
SBLUP	$\beta \sim N\left(0, \frac{h_g^2}{m}\right)$ h_g^2: SNP-based heritability, m: number of SNPs; $\lambda = m(1 - h_g^2)/h_g^2$	No	λ LD radius in kb	–
Ldpred2-Inf	Same as SBLUP	No	h_g^2 LD radius in cM or kb	–
LDpred-funct	$\beta_j \sim N(0, c\sigma_j^2)$ $\sum_{j=1}^{M} 1_{\sigma_j^2>0} c\sigma_j^2 = h_g^2$, c is a normalizing constant, σ_j^2 is the expected per SNP heritability under the baseline-LD annotation model estimated by stratified LDSC from the discovery GWAS within LDpred-funct software	No	h_g^2 LD radius in number of SNPs	–
LDpred2	$\beta_j \sim \begin{cases} N\left(0, \frac{h_g^2}{\pi m}\right), & \text{with probability of } \pi \\ 0, & \text{with probability of } 1-\pi \end{cases}$ When sparsity is "true," the β_j for SNPs in the $(1 - \pi)$ partition are all set to zero	Yes	h_g^2 π software default values, LD radius in cM or kb	π, sparsity
Lassosum	$f(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{y} + (1 - s)\boldsymbol{\beta}^T\mathbf{X}_r^T\mathbf{X}_r\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + s\boldsymbol{\beta}^T\boldsymbol{\beta} + 2\lambda\|\boldsymbol{\beta}\|_1^1$ $\mathbf{X}_r$: $n \times m$ matrix of genotypes of LD reference sample, where n is sample size	Yes	LD blocks	λ, s
PRS-CS	$\beta_j \sim N\left(0, \frac{\sigma^2}{n}\psi_j\right)$ $\psi_j \sim G(a, \delta_j)$ $\delta_j \sim G(b, \phi)$, ϕ is a global scaling parameter	Yes	$a = 1$, $b = 0.5$ n LD blocks	ϕ
PRS-CS-auto	Same as PRS-CS, but estimates ϕ from the discovery GWAS	No	$a = 1$, $b = 0.5$ n LD blocks	–
SBayesR	$\beta_j \mid \boldsymbol{\pi}, \sigma_\beta^2 \sim \begin{cases} 0, & \text{with probability of } \pi_1 \\ N(0, \gamma_2\sigma_\beta^2), & \text{with probability of } \pi_2 \\ \vdots & \\ N(0, \gamma_c\sigma_\beta^2), & \text{with probability of } 1 - \sum_{c=1}^{C-1}\pi_c \end{cases}$ $\sigma_\beta^2 \sim Inv - \chi^2(d.f. = 4)$ $\pi_i \sim Dir(\mathbf{1})$, estimated from discovery GWAS in SBayesR software γ_i are scaling parameters	No	LD radius in cM or kb C = 4 γ software default values	–
MegaPRS	Lasso: $\beta_j \sim DE(\lambda/\sigma_j)$ Ridge regression: $\beta_j \sim N(0, v\sigma_j^2)$ BOLT-LMM: $\beta_j \sim \begin{cases} N\left(0, \frac{(1-f_2)\sigma_j^2}{\pi}\right), & \text{with probability of } \pi \\ N\left(0, \frac{f_2\sigma_j^2}{1-\pi}\right), & \text{with probability of } 1-\pi \end{cases}$ f_2 is the proportion of the total mixture variance in the second normal distribution BayesR: similar to SBayesR with C = 4, and π_i and γ_i estimated in the tuning sample σ_j^2 is the expected per SNP-heritability under BLD-LDAK model using SumHer	Yes	LD radius in cM or kb Parameters used in BLD-LDAK Grid search parameter values for each method	The tuning cohort is used to estimate the parameters that maximize prediction for each model, and from these the model that maximizes prediction is selected

Distributions: *N*: normal distribution; *G*: gamma distribution; *Inv* $- \chi^2$: inverse chi-squared distribution, *Dir*: Dirichlet distribution; *DE*: double exponential distribution; $\|\boldsymbol{\beta}\|_1^1 = \sum_i|\beta_i|$. When h_g^2 (SNP-based heritability) is a predefined parameter, it is estimated from the discovery GWAS, where discovery GWAS is the genome-wide set of association statistics (SNP identification number, reference allele, frequency of reference allele, association effect size for reference allele, standard error of effect size, association *p* value, sample size). Bold indicates matrix notation, and italic indicates scalar notation. All methods require a reference sample with genotypes to model LD between SNPs.

cM, centimorgan; GWAS, genome-wide association study; kb, kilobase pair; LD, linkage disequilibrium; SNP, single nucleotide polymorphism.

are selected for application in the target sample. Several methods employ an LD reference sample to infer the expected correlation structure between SNP association statistics; those recommended by each software implementation were used.

Evaluation of Out-Of-Sample Prediction

The accuracy of prediction in each target cohort was quantified by the following statistics:

1. Area under the receiver operator characteristic curve (AUC) [R library pROC (28)]. AUC can be interpreted as a probability that a case ranks higher than a control.
2. The proportion of variance on the liability scale explained by PGS (29). We used the population lifetime risk of SCZ and MDD as 1% and 15%, respectively, to convert the variance explained in a linear regression to the liability scale (24,30,31).
3. Odds ratio (OR) of tenth PGS decile relative to the first decile.
4. OR of tenth PGS decile relative to those ranked in the middle of the PGS distribution, which is calculated as the average of OR of tenth decile relative to fifth and sixth decile.
5. Standard deviation unit increase in cases. The PGSs in each target cohort were scaled by standardizing the PGSs of controls and applying the standardization to cases: $\frac{PGS_{case} - mean(PGS_{control})}{SD(PGS_{control})}$, where SD is standard deviation. This does not impact PGS evaluation statistics but simply means that PGSs are in standard deviation units for all cohorts.

The regression analyses for evaluation statistics 2 through 4 include 6 ancestry principal components as covariates. These covariates are not included in the AUC model and the standard deviation unit increase in cases model (see Supplement 1).

RESULTS

Prediction evaluation statistics based on all 10 PGS methods and applied to SCZ across 30 study cohorts (Figure 1, Figure S1 in Supplement 1, and Tables S3 and S4 in Supplement 2) and to MDD across 26 cohorts (Figure S2 in Supplement 1 and Tables S5 and S6 in Supplement 2) are presented. There was variability in prediction statistics across target cohorts [as observed before (12,30)] that was not a reflection of sample size (Figure S3 in Supplement 1 and Table S4 in Supplement 2 for SCZ; Figure S4 in Supplement 1 and Table S6 in Supplement 2 for MDD). Some significant associations were found from regression of prediction statistics on principal components estimated from genome-wide SNPs for SCZ (Figure S3 in Supplement 1), but not MDD (Figure S4 in Supplement 1), where the principal components captured both within-European ancestry and array differences between cohorts. The correlations of PGS between different methods were high (Table S7 in Supplement 2), but were lowest between PC+T and other methods (minimum 0.68). In contrast, the correlations between the other 9 methods were always $>$ 0.82. In theory, LDpred2-Inf and SBLUP are the same method. In practice, there were differences in implementation (e.g., different input parameters associated with definition of LD window), and although the correlation between their PGSs was 0.974, the prediction accuracy was consistently higher for LDpred2-Inf. For SCZ, the AUC from methods that directly model genetic architecture, other than PRS-CS-auto, was significantly higher than the PC+T method at the nominal level (Figure 1A). PGSs from LDpred2, SBayesR, and MegaPRS were significantly higher than the PC+T method after Bonferroni correction ($p < .0011 = .05/45$ (45 pairwise comparisons between 10 methods), one-tailed Student's t test). For MDD, none of the differences between methods were significant (Figure S2A in Supplement 1). For both SCZ and MDD across all statistics, regardless of tuning cohorts, LDpred2, SBayesR, and MegaPRS showed relatively better performance (median across target cohorts) than other methods, although there was no significant difference between the 9 methods that directly model genetic architecture. For variance explained on the liability scale, the PC+T PGS explained 6.4% for SCZ, averaged over the median values across the 4 tuning cohorts (Figure 1B), while this variance was 8.9%, 9.0%, and 9.2% for MegaPRS, LDpred2, and SBayesR, corresponding to an increase of 39%, 41%, and 44%, respectively. For MDD, although the variance explained is lower in absolute terms, 2.2% for PC+T versus 3.4% for MegaPRS, 3.5% for LDpred2, and 3.5% for SBayesR, the variance of SBayesR represents a 59% increase (Figure S2B in Supplement 1).

We provide several evaluation statistics that focus on those in the top 10% of PGSs because clinical utility of PGSs for psychiatric disorders is likely to focus on individuals who are in the top tail of the distribution of predicted genetic risk. The ORs for top versus bottom decile were large, ranging from 14 for PC+T to 30 for MegaPRS for SCZ and 3 for PC+T to 3.7 for SBayesR for MDD. While these top versus bottom decile ORs (Figure 1C and Figure S2C in Supplement 1) were much larger than the OR obtained by using PGSs to screen a general population (Figure 1D and Figure S2D in Supplement 1) or patients in a health care system to identify people at high risk (32,33), these comparisons are useful for research purposes, which could, for example, make experimental designs focusing on individuals with high versus low PGSs cost-effective (34). The ORs of top 10% versus middle 10% were much less impressive, up to median of 6 for SCZ and 2 for MDD, but more fairly represent the value of PGSs in population settings. These values can be benchmarked against risk in first-degree relatives of affected individuals, which are on the order of 8 for SCZ and 2 for MDD; low values are always expected for MDD because it is more common (lifetime risk approximately 15% compared with approximately 1% for SCZ). The ORs were particularly high for some cohorts (Table S4 in Supplement 2) because in some SCZ cohorts the bottom 10% included very few or no cases, especially in cohorts with relatively small sample sizes.

Impact of Tuning Cohort

Five methods (i.e., PC+T, LDpred2, Lassosum, PRC-CS, and MegaPRS) use tuning cohorts to determine key parameters for application of the method in the target cohorts. Tuning parameters impact results in two ways. First, the parameters may be dependent on the choice of tuning cohort. Second, the discovery GWAS sample may be reduced in size (and hence

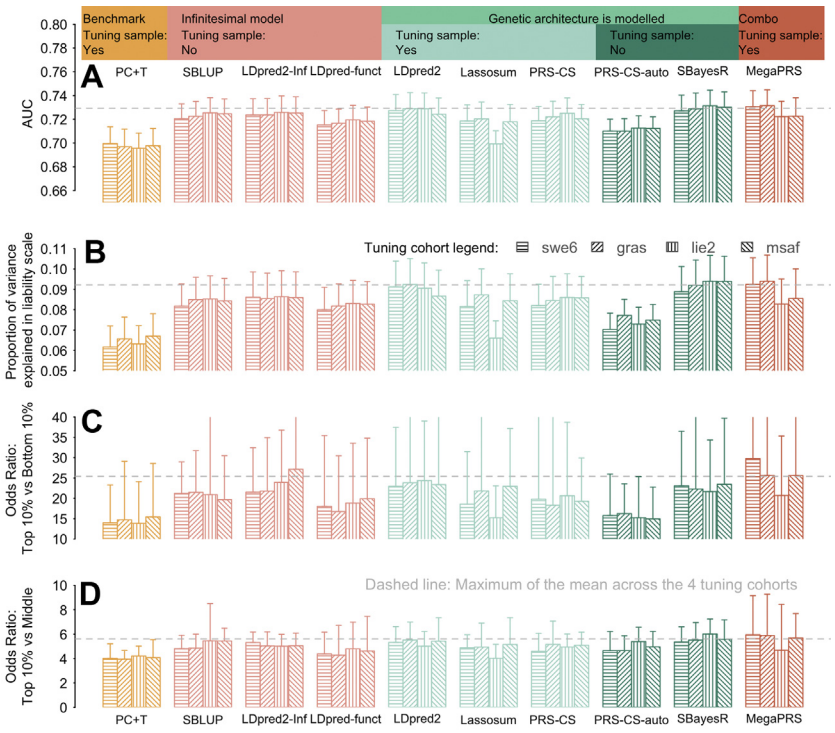

Figure 1. Prediction results for schizophrenia case/control status using different polygenic score (PGS) methods. The PGSs were constructed from schizophrenia genome-wide association study summary statistics excluding the target cohort and a tuning cohort (shading legend). Each bar reflects the median across 30 target cohorts; whiskers show the 95% confidence interval for comparing medians. **(A)** The area under the receiver operator characteristic curve (AUC) statistic can be interpreted as the probability that a case ranks higher than a control. **(B)** The proportion of variance explained by PGSs on the scale of liability, assuming a population lifetime risk of 1%. **(C)** The odds ratio when considering the odds of being a case comparing the top 10% vs. bottom 10% of PGSs. **(D)** The odds of being a case in the top 10% of PGSs vs. the odds of being a case in the middle of the PGS distribution. The middle was calculated as the averaged odds ratio of the top 10% ranked on PGSs relative to the 5th decile and 6th decile. PC+T (also known as P+T) is the benchmark method and is shown in orange. The methods that use an infinitesimal model assumption are shown in pink. The methods that model the genetic architecture are shown in green; light green shows the methods using a tuning cohort to determine the genetic architecture of a trait, and dark green shows the methods learning the genetic architecture from a discovery sample, without using a tuning cohort. MegaPRS using the BLD-LDAK model that assumes the distribution of single nucleotide polymorphism effect depends on its allele frequency, linkage disequilibrium, and function annotation is shown in dark orange. MegaPRS assigns 4 priors to each single nucleotide polymorphism: LASSO, Ridge, BOLT-LMM, BayesR. Each prior has different hyperparameters that are identified using the tuning cohort. The dashed gray lines are the maximum of the average across the 4 tuning cohorts. The sample sizes of the tuning cohorts are swe6, 1094 cases, 1219 controls; lie2, 137 cases, 269 controls; msaf, 327 cases, 139 controls; and gras, 1086 cases, 1232 controls.

power) if a tuning cohort needs to be excluded from the discovery GWAS. In all our analyses, the tuning cohort was excluded from all GWAS discovery samples so that the GWAS discovery sample was not variable across methods for each target cohort. Our results show that the tuning cohort can have considerable impact (Figures 1 and 2). In our results, the tuning cohort that generated higher PGSs was method dependent and differed between cohorts. For the methods that used tuning samples, the larger tuning samples (swe6 and gras) mostly generated higher prediction statistics compared with the two smaller tuning samples (lie and msaf), but the differences were not statistically significant. Although methods SBLUP, LDpred2-Inf, LDpred-funct, PRS-CS-auto, and SBayesR require no tuning cohort, they serve as a benchmark, as the differences in their results reflect differences in the changed discovery samples (e.g., msaf is in the discovery sample when swe6 is the tuning cohort, and vice versa) as well as the stochasticity inherent in the Gibbs sampling of Bayesian methods.

Impact of MAF/INFO Threshold

A MAF threshold of 0.1 and an INFO threshold of 0.9 were used to be consistent with applications in the PGC SCZ (30) and PGC MDD (24) studies, which had been imposed recognizing that these thresholds generated more robust PGS results than using lower threshold values. In the second sensitivity analysis applied to the SCZ data, the MAF threshold was relaxed to 0.05 or 0.01 (Figure 3). The prediction evaluation statistics increased for some cohorts and decreased for others (trends with sample size were not significant). PC+T was more impacted that the other 9 methods. Across target cohorts, different evaluation statistics were almost identical when including less common SNPs (Table S3 in Supplement 2). Relaxing the INFO score to 0.3 had a negligible effect (Figure S5 in Supplement 1).

DISCUSSION

Comparison of PGS risk prediction methods showed that all 9 methods that directly model genetic architecture had higher prediction evaluation statistics over the benchmark PC+T method for SCZ and MDD. While the differences between these 9 methods were small, we found that MegaPRS, LDpred2, and SBayesR consistently ranked highest. Given that the PGS is a sum of many small effects, a normal distribution of PGSs in a population is expected (and observed, as shown in Figures S6–S9 in Supplement 1). In idealized data, such as the relatively simple simulation scenarios usually considered in method development, all evaluation statistics should rank the methods in the same order, but with real data sets this is not guaranteed. This is the motivation for considering a range of evaluation statistics. Our focus on statistics for those in the top 10% of PGSs is relevant to potential clinical utility. In the context of psychiatry, it is likely that this will focus on people presenting in a prodromal state with clinical symptoms that have not yet been recognized to be specific to a diagnosis

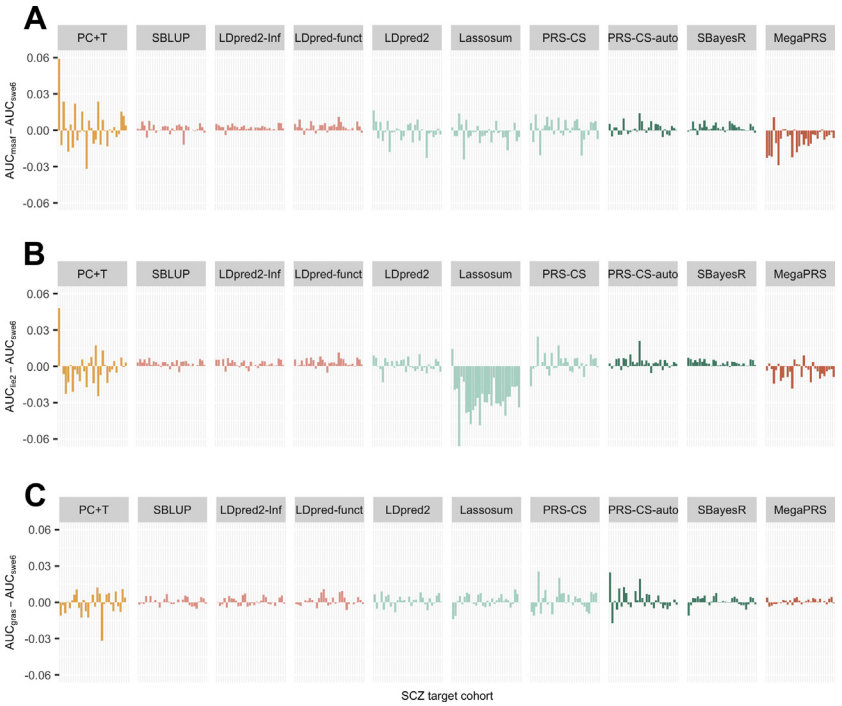

Figure 2. (A–C) Sensitivity analyses using different tuning cohorts comparing different polygenic score methods. Differences in the area under the receiver operator characteristic curve (AUC) of schizophrenia (SCZ) of a polygenic score method when using different tuning cohorts. The different bars in each method (x-axis) refer to different validation cohorts ordered by sample size. The y-axis is the AUC difference when using alternative tuning cohorts, i.e., lie2 (137 cases, 269 controls), msaf (327 cases, 139 controls), or gras (1086 cases, 1232 controls), compared with swe6 (1094 cases, 1219 controls). The minor allele frequency quality control threshold is 0.1. Note: SBLUP, LDpred2-Inf and LDpred-funct, PRS-CS-auto, and SBayesR do not need a tuning cohort, but serve as a benchmark to the other methods, which need a tuning cohort. These methods differ when a different tuning cohort is left out because the discovery genome-wide association study also changes.

(11,35). High PGSs in people presenting to clinics could contribute to clinical decision making by identifying individuals for closer monitoring or earlier intervention. As a genetic-based predictor predicts only part of the risk of disease, and as a PGS predicts only part of the genetic contribution to disease, it is acknowledged that PGSs cannot be fully accurate predictors. Hence, the discriminative ability of PGSs is low in the general population, and the use of PGSs in clinical settings requires evaluation, including related ethical issues (4). Nonetheless, PGSs in combination with clinical risk factors could make a useful contribution to risk prediction (35–37).

In sensitivity analyses that used different quality criteria for SNPs, e.g., MAF of 0.01 versus 0.05 or INFO of 0.3 versus 0.9, we concluded that currently there is little to be gained in PGSs from including SNPs with MAF $<$ 0.10 and INFO $<$ 0.9 for the diseases/data set studied (Tables S8 and S9 in Supplement 2). This result may seem counterintuitive, as variants with low MAF are expected to play an important role in common diseases, and some may be expected to have larger effect sizes than more common variants (38,39). However, sampling variance is a function of allele frequency, $\approx$ var (y)/[2 * MAF (1 − MAF) * n], where y is the phenotype and n is sample size, such that a variant of MAF = 0.01 has sampling variance 9 times greater than a variant of MAF = 0.1. Moreover, in real data sets, small sample size of contributing cohorts means that technical artifacts can accumulate to increase error in effect size

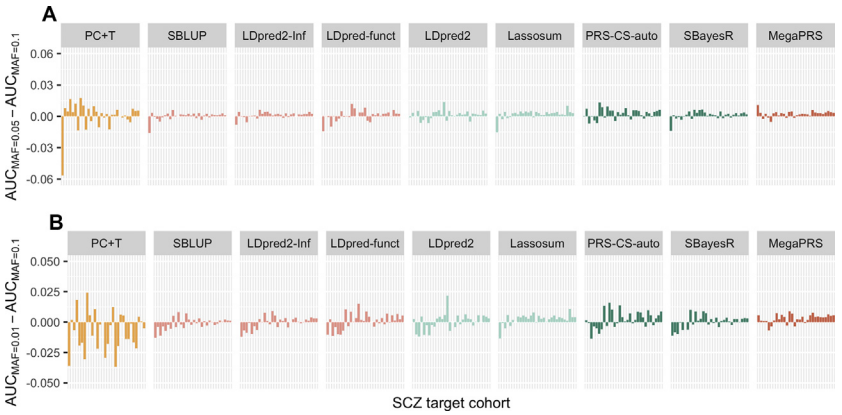

Figure 3. Sensitivity analyses using different minor allele frequency (MAF) quality control thresholds. Differences in area under the receiver operator characteristic curve (AUC) of schizophrenia (SCZ) of a polygenic score method when using different MAF quality control thresholds. The different bars in each method (x-axis) refer to different validation cohorts ordered by sample size. The y-axis is the AUC difference between analyses using **(A)** MAF $<$ 0.05 and MAF $<$ 0.1 and **(B)** MAF $<$ 0.01 and MAF $<$ 0.1 as a quality control threshold. The tuning cohort is swe6.

estimates, particularly of low-frequency variants. As larger individual cohorts in discovery samples accumulate, our conclusion that little is gained from including variants of MAF < 0.1 or from reducing INFO threshold will need to be revisited. Moreover, our comparison of methods uses only study samples of European ancestry. More research and data are needed to understand the properties of prediction methods within other ancestries and across ancestries, given potential differences in genetic architectures (in terms of number, frequencies, and effect sizes of causal variants) and LD between measured variants and causal variants (40,41).

For both SCZ and MDD, while the methods other than PC+T had similar performance, LDpred2, MegaPRS, and SBayesR saw the highest prediction accuracy in most of the comparisons. We note that we did not consider a version of PC+T that has been shown to have higher out-of-sample prediction compared with the standard implementation (13). This method conducts a grid search in a tuning cohort to determine LD r^2 and INFO score thresholds for SNPs as well as the p-value threshold. As the optimal LD threshold is likely to vary across genomic regions, the grid search approach is less appealing than the methods that implicitly allow this to vary. A sensitivity analysis in which we varied the r^2 threshold in the PC+T showed only a small gain from optimizing this (Table S10 in Supplement 2). LDpred2 has a version that does not require a tuning sample, LDpred2-auto, but the authors showed that the two methods give similar results. SBayesR assumes that the SNP effects are drawn from a mixture of 4 distributions, which allows more flexibility in distributions of SNP effects by varying the proportion of SNPs in each distribution. Hence, SBayesR can fit essentially any underlying architecture in terms of variance explained by each SNP so that the SBLUP, LDpred2-Inf, and LDpred2 models are, in principle, special cases of the mixture model used in SBayesR (although method implementations are different). In addition to traits with a highly polygenic genetic architecture, we have recently shown that SBayesR outperforms other methods for two less polygenic diseases, Alzheimer's disease (42) (which includes the *APOE* locus, which has a very large effect size) and amyotrophic lateral sclerosis (43) [for which there is evidence of greater importance of low MAF variants compared to SCZ (44)]. The original SBayesR publication showed that in both simulations and applications to real data, the method performed well across a range of traits with different underlying genetic architectures. MegaPRS uses 4 different priors for the distribution of SNP effect, i.e., Lasso, Ridge, BOLT-LMM, and BayesR (Table 1). It rescales SNP effects based on each of those priors and for each method selects the combination of parameters that maximizes prediction in the tuning sample and then selects the best method among these. Hence, MegaPRS is a collection of the other methods, and the SNP distribution selected varies depending on both tuning and target (Table S11 in Supplement 2). Here we found that it selects BayesR 87% of the time when the tuning samples are large (otherwise BOLT-LMM) and selects Lasso 78% of the time when tuning samples are small. We implemented MegaPRS using the BLD-LDAK model recommended by the authors, which assumes that the distribution of SNP effects depends on its allele frequency and functional annotation. While adding functional annotation to up or down weight SNPs is appealing, in practice there seemed to be no advantage in MegaPRS compared with LDpred2 and SBayesR, which did not use functional annotations. Surprisingly, LDpred-funct performed consistently less well than LDpred2-Inf, but this should be revisited, as currently the LDpred-funct article is available only as a preprint (16).

Another study has compared 8 PGS methods for 8 disease/disorder traits (including MDD) and 3 continuous phenotypes comparing methods in two large community samples, the UK Biobank and the Twins Early Development Study (45). Consistent with our results, SBayesR attained a high prediction accuracy for MDD, although performance of SBayesR was reported to vary across traits. As SBayesR expects effect size estimates and their standard errors to have properties consistent with the sample size and with the LD patterns imposed from an external reference panel, if GWAS summary statistics have nonideal properties (perhaps resulting from meta-analysis errors or approximations), SBayesR may not achieve converged solutions. SBayesR in general is more sensitive to any inconsistent properties between GWAS and LD reference samples than methods that select hyperparameters based on cross-validation in a tuning sample, such as LDpred2 (15). We note that the above-mentioned LDpred-funct preprint article reported that SBayesR performed well across a range of quantitative and binary traits. A key advantage of SBayesR is that there is no need for the user to tune or select model or software parameters. Moreover, it does not need a tuning cohort to derive SNP effect weights, but rather learns the genetic architecture from the properties of the GWAS results. Computationally, SBayesR is also very efficient; using one central processing unit (CPU), it takes approximately 2 hours to generate SNP weights based on each discovery sample and predict into the left-out-cohort using a Markov Chain Monte Carlo chain of 10,000 iterations (the computing time can be reduced by running a shorter chain as a negligible change in prediction accuracy was found after 4000 iterations), which compares to PRS-CS, 40 hours using 5 CPUs; LDpred2, 5 hours using 15 CPUs; and MegaPRS, 1 hour using 5 CPUs. Last, given that SBayesR uses only HapMap3 SNPs that are mostly well imputed, it should be possible to provide these SBayesR SNP weights as part of a GWAS pipeline to apply in external target samples.

All methods are compared using their default parameter settings. An optimal setting of each method could potentially increase the prediction accuracy. Most likely the optimal parameter settings are trait (genetic architecture) dependent (13). In this study, we found that all methods that more formally model the genetic architecture than PC+T perform better than PC+T, but there is little to choose between those methods. For application in psychiatric disorders, which are all highly polygenic traits, we particularly recommend LDpred2, MegaPRS, and SBayesR, which consistently rank high in all comparisons.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by the National Health and Medical Research Council (Grant Nos. 1173790, 1078901, and 108788 [to NRW] and Grant No. 1113400 [to NRW and PMV]) and the Australian Research Council (Grant No. FL180100072 [to PMV]).

This work would not have been possible without the contributions of the investigators who comprise the PGC SCZ and PGC MDD Working Groups.

For a full list of acknowledgments of all individual cohorts included in PGC SCZ and PGCMDD Working Groups, please see the original publications. The PGC has received major funding from the National Institute of Mental Health (Grant No. U01 MH109528).

The Münster cohort was funded by the German Research Foundation (Grant No. FOR2107 DA1151/5-1 and DA1151/5-2 [to Udo Dannlowski] and Grant No. SFB-TRR58, Projects C09 and Z02 [to Udo Dannlowski]) and Interdisciplinary Center for Clinical Research of the Faculty of Medicine of Münster (Grant No. Dan3/012/17 [to Udo Dannlowski]).

Some data used in this study were obtained from the database of Genotypes and Phenotypes (dbGaP). dbGaP Study Accession phs000021: Funding support for the Genome-Wide Association of Schizophrenia Study was provided by the National Institute of Mental Health (Grant Nos. R01 MH67257, R01 MH59588, R01 MH59571, R01 MH59565, R01 MH59587, R01 MH60870, R01 MH59566, R01 MH59586, R01 MH61675, R01 MH60879, R01 MH81800, U01 MH46276, U01 MH46289, U01 MH46318, U01 MH79469, and U01 MH79470), and the genotyping of samples was provided through the Genetic Association Information Network. Samples and associated phenotype data for the Genome-Wide Association of Schizophrenia Study were provided by the Molecular Genetics of Schizophrenia Collaboration (principal investigator P.V. Gejman, Evanston Northwestern Healthcare and Northwestern University, Evanston, IL). dbGaP accession phs000196: This work used in part data from the National Institute of Neurological Disorders and Stroke dbGaP database from the Center for Inherited Disease Research:NeuroGenetics Research Consortium Parkinson's Disease Study. dbGaP accession phs000187: High-Density SNP Association Analysis of Melanoma: Case-Control and Outcomes Investigation. Research support to collect data and develop an application to support this project was provided by the National Institutes of Health (Grant Nos. P50 CA093459, P50 CA097007, R01 ES011740, and R01 CA133996).

Statistical analyses were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org) hosted by SURFsara and financially supported by the Netherlands Scientific Organization (Grant No. 480-05-003) along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam.

We thank the customers, research participants, and employees of 23andMe for making this work possible. The study protocol used by 23andMe was approved by an external Association for the Accreditation of Human Research Protection Programs–accredited institutional review board.

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

ARTICLE INFORMATION

From the Institute for Molecular Bioscience (GN, JZ, JAR, YW, ZZ, RR, JK, JY, PMV, NRW) and Queensland Brain Institute (NRW), University of Queensland; Faculty of Health (DRN), School of Biomedical Sciences, Centre for Genomics and Personalised Health, Queensland University of Technology, Brisbane, Queensland, Australia; Psychiatric and Neurodevelopmental Genetics Unit (TG, JWS) and Department of Psychiatry (JWS), Massachusetts General Hospital, Boston; Stanley Center for Psychiatric Research (JWS), Broad Institute, Cambridge, Massachusetts; Social, Genetic and Developmental Psychiatry Centre (JRIC), Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom; and School of Life Sciences (JY), Westlake University, Hangzhou, Zhejiang, China

The PGC SCZ Working Group is a collaborative coauthor of this article. The individual authors are Stephan Ripke, Benjamin M. Neale, Aiden Corvin, James T.R. Walters, Kai-How Farh, Peter A. Holmans, Phil Lee, Brendan Bulik-Sullivan, David A. Collier, Hailiang Huang, Tune H. Pers, Ingrid Agartz, Esben Agerbo, Margot Albus, Madeline Alexander, Farooq Amin, Silviu A. Bacanu, Martin Begemann, Richard A. Belliveau Jr., Judit Bene, Sarah E. Bergen, Elizabeth Bevilacqua, Tim B. Bigdeli, Donald W. Black, Richard Bruggeman, Nancy G. Buccola, Randy L. Buckner, William Byerley, Wiepke Cahn, Guiqing Cai, Dominique Campion, Rita M. Cantor, Vaughan J. Carr, Noa Carrera, Stanley V. Catts, Kimberley D. Chambert, Raymond C.K. Chan, Ronald Y.L. Chen, Eric Y.H. Chen, Wei Cheng, Eric F.C. Cheung, Siow Ann Chong, C. Robert Cloninger, David Cohen, Nadine Cohen, Paul Cormican, Nick Craddock, James J. Crowley, Michael Davidson, Kenneth L. Davis, Franziska Degenhardt, Jurgen Del Favero, Ditte Demontis, Dimitris Dikeos, Timothy Dinan, Srdjan Djurovic, Gary Donohoe, Elodie Drapeau, Jubao Duan, Frank Dudbridge, Naser Durmishi, Peter Eichhammer, Johan Eriksson, Valentina Escott-Price, Laurent Essioux, Ayman H. Fanous, Martilias S. Farrell, Josef Frank, Lude Franke, Robert Freedman, Nelson B. Freimer, Marion Friedl, Joseph I. Friedman, Menachem Fromer, Giulio Genovese, Lyudmila Georgieva, Ina Giegling, Paola Giusti-Rodríguez, Stephanie Godard, Jacqueline I. Goldstein, Vera Golimbet, Srihari Gopal, Jacob Gratten, Lieuwe de Haan, Christian Hammer, Marian L. Hamshere, Mark Hansen, Thomas Hansen, Vahram Haroutunian, Annette M. Hartmann, Frans A. Henskens, Stefan Herms, Joel N. Hirschhorn, Per Hoffmann, Andrea Hofman, Mads V. Hollegaard, David M. Hougaard, Masashi Ikeda, Inge Joa, Antonio Julià, René S. Kahn, Luba Kalaydjieva, Sena Karachanak-Yankova, Juha Karjalainen, David Kavanagh, Matthew C. Keller, James L. Kennedy, Andrey Khrunin, Yunjung Kim, Janis Klovins, James A. Knowles, Bettina Konte, Vaidutis Kucinskas, Zita Ausrele Kucinskiene, Hana Kuzelova-Ptackova, Anna K. Kähler, Claudine Laurent, Jimmy Lee, S. Hong Lee, Sophie E. Legge, Bernard Lerer, Miaoxin Li, Tao Li, Kung-Yee Liang, Jeffrey Lieberman, Svetlana Limborska, Carmel M. Loughland, Jan Lubinski, Jouko Lönnqvist, Milan Macek, Patrik K.E. Magnusson, Brion S. Maher, Wolfgang Maier, Jacques Mallet, Sara Marsal, Manuel Mattheisen, Morten Mattingsdal, Robert W. McCarley, Colm McDonald, Andrew M. McIntosh, Sandra Meier, Carin J. Meijer, Bela Melegh, Ingrid Melle, Raquelle I. Mesholam-Gately, Andres Metspalu, Patricia T. Michie, Lili Milani, Vihra Milanova, Younes Mokrab, Derek W. Morris, Ole Mors, Kieran C. Murphy, Robin M. Murray, Inez Myin-Germeys, Bertram Müller-Myhsok, Mari Nelis, Igor Nenadic, Deborah A. Nertney, Gerald Nestadt, Kristin K. Nicodemus, Liene Nikitina-Zake, Laura Nisenbaum, Annelie Nordin, Eadbhard O'Callaghan, Colm O'Dushlaine, F. Anthony O'Neill, Sang-Yun Oh, Ann Olincy, Line Olsen, Jim Van Os, Psychosis Endophenotypes International Consortium, Christos Pantelis, George N. Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T. Pato, Tiina Paunio, Milica Pejovic-Milovancevic, Diana O. Perkins, Olli Pietiläinen, Jonathan Pimm, Andrew J. Pocklington, John Powell, Alkes Price, Ann E. Pulver, Shaun M. Purcell, Digby Quested, Henrik B. Rasmussen, Abraham Reichenberg, Mark A. Reimers, Alexander L. Richards, Joshua L. Roffman, Panos Roussos, Douglas M. Ruderfer, Veikko Salomaa, Alan R. Sanders, Ulrich Schall, Christian R. Schubert, Thomas G. Schulze, Sibylle G. Schwab, Edward M. Scolnick, Rodney J. Scott, Larry J. Seidman, Jianxin Shi, Engilbert Sigurdsson, Teimuraz Silagadze, Jeremy M. Silverman, Kang Sim, Petr Slominsky, Jordan W. Smoller, Hon-Cheong So, Chris C.A. Spencer, Eli A. Stahl, Hreinn Stefansson, Stacy Steinberg, Elisabeth Stogmann, Richard E. Straub, Eric Strengman, Jana Strohmaier, T. Scott Stroup, Mythily Subramaniam, Jaana Suvisaari, Dragan M. Svrakic, Jin P. Szatkiewicz, Erik Söderman, Srinivas Thirumalai, Draga Toncheva, Sarah Tosato, Juha Veijola, John Waddington, Dermot Walsh, Dai Wang, Qiang Wang, Bradley T. Webb, Mark Weiser, Dieter B. Wildenauer, Nigel M. Williams, Stephanie Williams, Stephanie H. Witt, Aaron R. Wolen, Emily H.M. Wong, Brandon K. Wormley, Hualin Simon Xi, Clement C. Zai, Xuebin Zheng, Fritz Zimprich, Naomi R. Wray, Kari Stefansson, Peter M. Visscher, Wellcome Trust Case-Control Consortium, Rolf Adolfsson, Ole A. Andreassen, Douglas H.R. Blackwood, Elvira Bramon, Joseph D. Buxbaum, Anders D. Børglum, Sven Cichon, Ariel Darvasi, Enrico Domenici, Hannelore Ehrenreich, Tõnu Esko, Pablo V. Gejman, Michael Gill, Hugh Gurling, Christina M. Hultman, Nakao Iwata, Assen V. Jablensky, Erik G. Jönsson, Kenneth S. Kendler, George Kirov, Jo Knight, Todd Lencz, Douglas F. Levinson, Qingqin S. Li, Jianjun Liu, Anil K. Malhotra, Steven A. McCarroll, Andrew McQuillin, Jennifer L. Moran, Preben B. Mortensen, Bryan J. Mowry, Markus M. Nöthen, Roel A. Ophoff, Michael J. Owen, Aarno Palotie, Carlos N. Pato, Tracey L. Petryshen, Danielle Posthuma, Marcella Rietschel, Brien P. Riley, Dan Rujescu, Pak C. Sham, Pamela Sklar, David St Clair, Daniel R. Weinberger, Jens R. Wendland, Thomas Werge, Mark J. Daly, Patrick F. Sullivan, and Michael C. O'Donovan. (Affiliations are listed in Supplement 1.)

The PGC MDD Working Group is a collaborative coauthor of this article. The individual authors are Naomi R. Wray, Stephan Ripke, Manuel Mattheisen, Maciej Trzaskowski, Enda M. Byrne, Abdel Abdellaoui, Mark J. Adams, Esben Agerbo, Tracy M. Air, Till F.M. Andlauer, Silviu-Alin Bacanu, Marie Bækvad-Hansen, Aartjan T.F. Beekman, Tim B. Bigdeli, Elisabeth B.

Binder, Julien Bryois, Henriette N. Buttenschøn, Jonas Bybjerg-Grauholm, Na Cai, Enrique Castelao, Jane Hvarregaard Christensen, Toni-Kim Clarke, Jonathan R.I. Coleman, Lucía Colodro-Conde, Baptiste Couvy-Duchesne, Nick Craddock, Gregory E. Crawford, Gail Davies, Ian J. Deary, Franziska Degenhardt, Eske M. Derks, Nese Direk, Conor V. Dolan, Erin C. Dunn, Thalia C. Eley, Valentina Escott-Price, Farnush Farhadi Hassan Kiadeh, Hilary K. Finucane, Jerome C. Foo, Andreas J. Forstner, Josef Frank, Héléna A. Gaspar, Michael Gill, Fernando S. Goes, Scott D. Gordon, Jakob Grove, Lynsey S. Hall, Christine Søholm Hansen, Thomas F. Hansen, Stefan Herms, Ian B. Hickie, Per Hoffmann, Georg Homuth, Carsten Horn, Jouke-Jan Hottenga, David M. Hougaard, David M. Howard, Marcus Ising, Rick Jansen, Ian Jones, Lisa A. Jones, Eric Jorgenson, James A. Knowles, Isaac S. Kohane, Julia Kraft, Warren W. Kretzschmar, Zoltán Kutalik, Yihan Li, Penelope A. Lind, Donald J. MacIntyre, Dean F. MacKinnon, Robert M. Maier, Wolfgang Maier, Jonathan Marchini, Hamdi Mbarek, Patrick McGrath, Peter McGuffin, Sarah E. Medland, Divya Mehta, Christel M. Middeldorp, Evelin Mihailov, Yuri Milaneschi, Lili Milani, Francis M. Mondimore, Grant W. Montgomery, Sara Mostafavi, Niamh Mullins, Matthias Nauck, Bernard Ng, Michel G. Nivard, Dale R. Nyholt, Paul F. O'Reilly, Hogni Oskarsson, Michael J. Owen, Jodie N. Painter, Carsten Bøcker Pedersen, Marianne Giørtz Pedersen, Roseann E. Peterson, Wouter J. Peyrot, Giorgio Pistis, Danielle Posthuma, Jorge A. Quiroz, Per Qvist, John P. Rice, Brien P. Riley, Margarita Rivera, Saira Saeed Mirza, Robert Schoevers, Eva C. Schulte, Ling Shen, Jianxin Shi, Stanley I. Shyn, Engilbert Sigurdsson, Grant C.B. Sinnamon, Johannes H. Smit, Daniel J. Smith, Hreinn Stefansson, Stacy Steinberg, Fabian Streit, Jana Strohmaier, Katherine E. Tansey, Henning Teismann, Alexander Teumer, Wesley Thompson, Pippa A. Thomson, Thorgeir E. Thorgeirsson, Matthew Traylor, Jens Treutlein, Vassily Trubetskoy, André G. Uitterlinden, Daniel Umbricht, Sandra Van der Auwera, Albert M. van Hemert, Alexander Viktorin, Peter M. Visscher, Yunpeng Wang, Bradley T. Webb, Shantel Marie Weinsheimer, Jürgen Wellmann, Gonneke Willemsen, Stephanie H. Witt, Yang Wu, Hualin S. Xi, Jian Yang, Futao Zhang, Volker Arolt, Bernhard T. Baune, Klaus Berger, Dorret I. Boomsma, Sven Cichon, Udo Dannlowski, E.J.C. de Geus, J. Raymond DePaulo, Enrico Domenici, Katharina Domschke, Tõnu Esko, Hans J. Grabe, Steven P. Hamilton, Caroline Hayward, Andrew C. Heath, Kenneth S. Kendler, Stefan Kloiber, Glyn Lewis, Qingqin S. Li, Susanne Lucae, Pamela A.F. Madden, Patrik K. Magnusson, Nicholas G. Martin, Andrew M. McIntosh, Andres Metspalu, Ole Mors, Preben Bo Mortensen, Bertram Müller-Myhsok, Merete Nordentoft, Markus M. Nöthen, Michael C. O'Donovan, Sara A. Paciga, and Nancy L. Pedersen. (Affiliations are listed in Supplement 1.)

Address correspondence to Naomi R. Wray, Ph.D., at naomi.wray@uq.edu.au.

Received Sep 15, 2020; revised and accepted Apr 26, 2021.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.biopsych.2021.04.018.

REFERENCES

1. International Schizophrenia Consortium (2009): Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748–752.
2. Palk AC, Dalvie S, De Vries J, Martin AR, Stein DJ (2019): Potential use of clinical polygenic risk scores in psychiatry—ethical implications and communicating high polygenic risk. Philos Ethics Humanit Med 14:4.
3. Wray NR, Goddard ME, Visscher PM (2007): Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 17:1520–1528.
4. Wray NR, Lin T, Austin J, McGrath JJ, Hickie IB, Murray GK, *et al.* (2021): From basic science to clinical application of polygenic risk scores: A primer. JAMA Psychiatry 78:101–109.
5. Jenkins MA, Win AK, Dowty JG, MacInnis RJ, Makalic E, Schmidt DF, *et al.* (2019): Ability of known susceptibility snps to predict colorectal cancer risk for persons with and without a family history. Fam Cancer 18:389–397.
6. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, *et al.* (2019): Boadicea: A comprehensive breast cancer risk prediction model incorporating genetic and non-genetic risk factors. Genet Med 21:1708–1718.
7. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, *et al.* (2018): Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 50:1219–1224.
8. Lloyd-Jones DM, Wilson PWF, Larson MG, Beiser A, Leip EP, D'Agostino RB, *et al.* (2004): Framingham risk score and prediction of lifetime risk for coronary heart disease. Am J Cardiol 94: 20–24.
9. McCarthy MI, Mahajan A (2018): The value of genetic risk scores in precision medicine for diabetes. Expert Review of Precision Medicine and Drug Development 3:279–281.
10. Torkamani A, Wineinger NE, Topol EJ (2018): The personal and clinical utility of polygenic risk scores. Nat Rev Genet 19:581.
11. Murray GK, Lin T, Austin J, McGrath JJ, Hickie IB, Wray NR (2021): Could polygenic risk scores be useful in psychiatry? A review. JAMA Psychiatry 78:210–219.
12. Trzaskowski M, Mehta D, Peyrot WJ, Hawkes D, Davies D, Howard DM, *et al.* (2019): Quantifying between-cohort and between-sex genetic heterogeneity in major depressive disorder. Am J Med Genet B Neuropsychiatr Genet 180:439–447.
13. Privé F, Vilhjálmsson BJ, Aschard H, Blum MGB (2019): Making the most of clumping and thresholding for polygenic scores. Am J Hum Genet 105:1213–1221.
14. Robinson MR, Kleinman A, Graff M, Vinkhuyzen AAE, Couper D, Miller MB, *et al.* (2017): Genetic evidence of assortative mating in humans. Nat Hum Behav 1:0016.
15. Privé F, Arbel J, Vilhjálmsson BJ (2020): Ldpred2: Better, faster, stronger. Bioinformatics 36:5424–5431.
16. Márquez-Luna C, Gazal S, Loh PR, Kim SS, Furlotte N (2020): Ldpred-funct: Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andme data sets. bioRxiv. https://doi.org/10.1101/375337.
17. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC (2017): Polygenic scores via penalized regression on summary statistics. Genet Epidemiol 41:469–480.
18. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW (2019): Polygenic prediction via bayesian regression and continuous shrinkage priors. Nat Commun 10:1776.
19. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, *et al.* (2019): Improved polygenic prediction by bayesian multiple regression on summary statistics. Nat Commun 10:1–11.
20. Zhang Q, Prive F, Vilhjalmsson BJ, Speed D (2020): Improved genetic prediction of complex traits from individual-level data or summary statistics. bioRxiv. https://doi.org/10.1101/2020.08.24.265280.
21. Chatterjee N, Shi J, García-Closas M (2016): Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet 17:392.
22. Speed D, Balding DJ (2019): Sumher better estimates the SNP heritability of complex traits from summary statistics. Nat Genet 51:277–284.
23. Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, *et al.* (2018): Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet 50:381–389.
24. 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.
25. Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M, *et al.* (2019): Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci 22:343.
26. Sullivan PF, Geschwind DH (2019): Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. Cell 177:162–183.

27. The International HapMap 3 Consortium (2010): Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58.
28. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, *et al.* (2011): Proc: An open-source package for r and s+ to analyze and compare roc curves. BMC Bioinformatics 12:1–8.
29. Lee SH, Goddard ME, Wray NR, Visscher PM (2012): A better coefficient of determination for genetic profile analysis. Genet Epidemiol 36:214–224.
30. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014): Biological insights from 108 schizophrenia-associated genetic loci. Nature 511:421–427.
31. Lee SH, Wray NR, Goddard ME, Visscher PM (2011): Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88:294–305.
32. Zheutlin AB, Dennis J, Karlsson Linnér R, Moscati A, Restrepo N, Straub P, *et al.* (2019): Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 106,160 patients across four health care systems. Am J Psychiatry 176:846–855.
33. Binder EB (2019): Polygenic risk scores in schizophrenia: Ready for the real world? Am J Psychiatry 176:783–784.
34. Dobrindt K, Zhang H, Das D, Abdollahi S, Prorok T, Ghosh S, *et al.* (2020): Publicly available hiPSC lines with extreme polygenic risk scores for modeling schizophrenia. Complex Psychiatry 6:68–82.
35. Perkins DO, Olde Loohuis L, Barbee J, Ford J, Jeffries CD, Addington J, *et al.* (2020): Polygenic risk score contribution to psychosis prediction in a target population of persons at clinical high risk. Am J Psychiatry 177:155–163.
36. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, *et al.* (2018): Genomic risk prediction of coronary artery disease in 480,000 adults: Implications for primary prevention. J Am Coll Cardiol 72:1883–1893.
37. Cannon TD, Yu C, Addington J, Bearden CE, Cadenhead KS, Cornblatt BA, *et al.* (2016): An individualized risk calculator for research in prodromal psychosis. Am J Psychiatry 173:980–988.
38. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, *et al.* (2011): Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci 108:18026–18031.
39. Bomba L, Walter K, Soranzo N (2017): The impact of rare and low-frequency genetic variants in common disease. Genome Biol 18:77.
40. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ (2019): Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 51:584.
41. Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, *et al.* (2019): Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. Cell 179:589–603.
42. Zhang Q, Sidorenko J, Couvy-Duchesne B, Marioni RE, Wright MJ, Goate AM, *et al.* (2020): Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. Nat Commun 11:4799.
43. Restuadi R, Garton FC, Benyamin B, Lin T, Williams KL, Vinkhuyzen A, *et al.* (2021): Polygenic risk score analysis for amyotrophic lateral sclerosis leveraging cognitive performance, educational attainment and schizophrenia [published online ahead of print Apr 27]. Eur J Hum Genet.
44. van Rheenen W, Shatunov A, Dekker AM, McLaughlin RL, Diekstra FP, Pulit SL, *et al.* (2016): Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat Genet 48:1043–1048.
45. Pain O, Glanville KP, Hagenaars SP, Selzam SP, Fürtjes AE, Gaspar HA, *et al.* (2021): Evaluation of polygenic prediction methodology within a reference-standardized framework. PLoS Genet 17: e1009021.