A comprehensive evaluation of polygenic score methods across cohorts in psychiatric disorders

Authors:

Guiyan Ni,[1] Jian Zeng,[1] Joana A Revez,[1] Ying Wang,[1] Tian Ge,[2] Restaudi Restaudi,[1] Jacqueline Kiewa,[1] Dale R Nyholt,[3] Jonathan R I Coleman,[4] Jordan W Smoller,[5,2,6] Schizophrenia Working Group of the Psychiatric Genomics Consortium,[7] Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium,[8] Jian Yang,[1] Peter M Visscher,[1] Naomi R Wray[1,9]

1. Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, 4072, Australia
2. Psychiatric and Neurodevelopmental Genetics Unit (PNGU), Massachusetts General Hospital, Boston, MA, 02114, US
3. Faculty of Health, School of Biomedical Sciences, Centre for Genomics and Personalised Health, Queensland University of Technology, Brisbane, Queensland, 4000, Australia
4. Social, Genetic, and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, SE58AF United Kingdom
5. Department of Psychiatry, Massachusetts General Hospital, Boston, MA, 02114, US
6. Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, 02142, US
7. A list of members and affiliations appears in the Supplementary Data.
8. A list of members and affiliations appears in the Supplementary Data.
9. Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, 4072, Australia

Correspondence:

*Naomi R Wray: naomi.wray@uq.edu.au

Abstract:

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 (GWASs). PGS methods differ in terms of which DNA variants are included in the score and the weights assigned to them. PGSs are evaluated in independent target samples of individuals with known disease status. Evaluation of new PGS methods are made using simulated data or single target cohort, however, in real data sets there can be heterogeneity between target sample cohorts, which could reflect a number of real or artefactual factors. The Psychiatric Genomics Consortium working groups for schizophrenia (SCZ) and major depressive disorder (MDD) bring together many independently collected case-control cohorts for GWAS meta-analysis. These resources are used here in repeated application of leave-one-cohort-out GWAS analyses, generating robust conclusions for PGS prediction applied across multiple target (left-out) cohorts. Eight PGS methods (P+T, SBLUP, LDpred-Inf, LDpred-funct, LDpred, PRS-CS, PRS-CS-auto, SBayesR) are compared. We found that SBayesR had the highest prediction evaluation statistics in most comparisons. For SCZ across 30 target cohorts, the SBayesR PGS achieved a mean area under the receiver operator characteristic curve (AUC) of 0.733, and explained 9.9% of variance on the liability scale. For MDD across 26 target cohorts, the AUC and variance explained were 0.601 and 4.0%, respectively. The variance explained by the SBayesR PGS was 46% and 43% higher for SCZ and MDD, respectively, compared to the basic p-value thresholding P+T method.

48 # Introduction

49

50 Polygenic scores (PGSs), which assess the genetic risk of individuals for a disease[1; 2], are
51 calculated as a weighted count of genetic risk alleles in the genome of an individual, with the
52 risk alleles and their weights typically derived from the results of genome-wide association
53 studies (GWAS).[3] PGS can be calculated for any trait or disease with sufficiently powered
54 GWAS ('discovery samples'). For many common complex genetic disorders, such as
55 cancers[4; 5] and heart disease[6; 7], there is increasing interest in trialling PGS for early disease
56 detection, prevention and intervention[8; 9]. In the context of psychiatric disorders, it has been
57 argued[10] that PGS may have utility in the context of youth mental health clinics, where young
58 people present with symptoms that have not yet crystallised to portray a clear treatment
59 pathway. A high PGS could nudge clinic decision making for those presenting in this
60 prodromal state.

61

62 There are now many methods to calculate PGSs, and the methods differ in terms of two key
63 criteria: which DNA variants to include (DNA variants here are limited to single nucleotide
64 polymorphisms, SNPs, but can include other DNA variants tested for association with a trait)
65 and what weights to allocate to them. While stringent thresholds are set to declare
66 significance for association of individual SNPs in GWAS, PGSs are robust to inclusion of
67 some false positives, and the maximum prediction from PGSs tested in target samples (i.e.,
68 GWAS samples independent of the GWAS discovery sample) may include nominally
69 associated SNPs. The optimum method to decide which SNPs to select and what weights to
70 allocate them, may differ between traits depending on the sample size of the discovery
71 GWAS and on the genetic architecture of the trait (the number, frequencies and effect sizes
72 of causal variants), particularly given the linkage disequilibrium (LD) structure between

73 SNPs. Often, when new PGS methods are introduced, comparisons are made between a
74 limited set of methods using simulated data, together with application to some real data
75 examples. However, it can be difficult to compare across the new methods, particularly
76 because in real data there can be heterogeneity in PGS evaluation statistics between target
77 samples, not encountered in idealised simulations. The reasons for this heterogeneity are
78 usually unknown but could reflect a number of factors such as phenotype definition,
79 ascertainment strategies of cases and controls, cohort-specific ancestry within the broad
80 classification of ancestry defined by the GWAS discovery samples (e.g., European), or
81 technical artefacts in genotype generation.
82
83 Here, we compare eight PGS methods (P+T[3; 11], SBLUP[12], LDpred-Inf[13], LDpred[13], LDpred-
84 funct[14], PRS-CS[15], PRS-CS-auto[15] and SBayesR[16] in **Table 1**). Some of these methods (P+T,
85 LDpred and PRS-CS) require a tuning sample, a GWAS cohort with known trait status that is
86 independent of both discovery and target samples, used to select parameters needed to
87 generate the PGSs in the target sample. Briefly, P+T (pruning with a p-value threshold) uses
88 the GWAS effect size estimates as SNP weights and includes independent SNPs (defined by
89 an LD r^2 filter for a given chromosomal window distance) with association p-values lower
90 than a threshold (chosen after application in a tuning sample). P+T is the most commonly
91 used and basic method, and so is the bench-mark method here. The other methods (referred to
92 here as recent methods) assume either that all SNPs have an effect size drawn from a normal
93 distribution (SBLUP and LDpred-Inf) or that SNP effects are drawn from mixtures of
94 distributions with the key parameters defining these architectures estimated through Bayesian
95 frameworks (LDpred, PRS-CS, SBayesR). LDpred-funct includes functional annotation to
96 SNPs to up/down weight their contributions to the PGSs, which could improve prediction
97 accuracy if this functional information helps to better separate true and false positive

98 associations[17]. We apply these methods to data from the Psychiatric Genomics Consortium
99 (PGC) working groups for schizophrenia (SCZ)[18; 19] and major depressive disorder (MDD)[20-
100 22] (**Tables S1 and S2**). The PGC provides a useful resource for undertaking this study
101 because it brings together many independently collected cohorts for GWAS meta-analysis.
102 This allows the application of repeated leave-one-cohort-out GWAS analyses generating
103 robust conclusions from evaluation of PGS applied across multiple left-out target cohorts.

104 Materials and Methods

105 Data:

106 Schizophrenia GWAS summary statistics, denoted as PGC-SCZ2+, were available from PGC
107 Schizophrenia (SCZ) Working group (34 European ancestry cohorts, denoted as SCZ34)[18]
108 and another three cohorts from Pardiñas et al[19]. PGC-SCZ2+ comprises more than 8M
109 imputed SNPs in 31K SCZ cases and 41K controls. Individual level genotype data were
110 available from 25K cases and 30K controls of SCZ34. Detailed information about the cohorts
111 is provided elsewhere[23] but is summarised in **Table S1**. Since some methods require a tuning
112 sample (defined below), we arbitrarily chose the lie2 cohort (137 cases and 269 controls) as
113 the tuning cohort. The GWAS discovery sample was a meta-analysis of 35 cohorts; lie2 was
114 always excluded and then each of the remaining 33 cohorts was left-out in turn and used as
115 the target sample. In sensitivity analyses, investigating the impact of the tuning sample, the
116 msaf, gras and swe6 cohort were exchanged with lie2 in turn, in which msaf has a similar
117 sample size as lie2, 327 cases and 139 controls, while gras and swe6 are larger with more
118 than 2000 individuals each.
119
120 Major depression GWAS summary statistics were available from seven studies including UK
121 Biobank[21; 24], 23andMe[25], GERA[26], iPSYCH[27], deCODE[28], GenScotland[29; 30], and the PGC

122 Major Depressive Disorder (MDD) Working group (with the data previously denoted as
123 PGC29, but here MDD29)[20]. All are European ancestry studies and comprise almost 13M
124 imputed SNPs from 248K cases and 563K controls. MDD29 includes the GWAS results from
125 29 research study cohorts. Detailed information of the MDD29 cohorts is described
126 elsewhere[20; 21; 25-30] but is summarised in **Table S2**. Individual level genotype data were
127 available for 15K cases and 24K controls from 26 cohorts. We left one cohort out of those 26
128 cohorts in turn as the target sample. A cohort from Muenster[20], not included in the MDD29
129 was used as the tuning sample (845 clinical defined MDD cases and 834 controls). We then
130 meta-analysed with the other GWAS summary statistics results to make the discovery
131 samples. We note that the discovery sample meta-analyses include samples where the
132 depression phenotype is self-reported rather than following a structured clinical interview,
133 nonetheless we refer to the prediction as MDD since the PGC target cohorts are of MDD
134 cases and controls. 959 overlapped individuals between UK Biobank and MDD29 were
135 excluded from the target cohorts.
136
137 The datasets stored in the PGC central server follow strict guidelines with local ethics
138 committee approval.
139
140 Baseline SNP selection
141 For baseline analyses, only SNPs with minor allele frequency (MAF) > 0.1 and imputation
142 INFO score > 0.9 (converted to best-guess genotype values of 0, 1 or 2) were selected.
143 Sensitivity analyses relaxed the MAF threshold to MAF > 0.05 or 0.01 and INFO score
144 threshold to 0.3. All methods were conducted using HapMap3 SNPs, except the method P+T,
145 which was conducted based on all imputed SNPs (8M in SCZ, and 13M in MDD).
146

147 Prediction methods

148 We define a PGS of an individual, j, as a weighted sum of SNP allele counts: $\sum_{i=1}^{m} \hat{b}_i x_{ij}$,
149 where m is the number of SNPs included in the predictor, $\hat{b}_i$ is the per allele weight for the
150 SNP, x_{ij} is a count of the number (0, 1, or 2) of trait-associated alleles of SNP i in individual
151 j. The cohort (target sample) for which PGSs are calculated is excluded from the meta-
152 analysis that generates the GWAS summary statistics (discovery sample), so that discovery
153 and target samples are independent. We compared eight risk prediction methods (detailed
154 below): The methods differ in terms of the SNPs selected for inclusion in the predictor and
155 the $\hat{b}_i$ values assigned to the SNPs. All methods use the GWAS summary statistics as the
156 starting point, but each makes choices differently for which SNPs to include and for the $\hat{b}_i$
157 values to assign. Briefly, the key differences between the methods are the assumptions made
158 about the underlying genetic architecture and the distributions of true effect sizes, with
159 Bayesian methods setting some priors for these distributions. Several methods employ an LD
160 reference sample to determine LD between SNPs. Here, we use EUR of the 1000 Genomes
161 Project as the LD reference, unless the method software provides an LD reference. In some
162 methods the PGS calculated in a target cohort requires estimates of parameter values, which
163 need to be estimated by application of the PGS method to a tuning cohort (also not included
164 in the discovery GWAS sample) using a range of parameter estimates, then selecting the
165 parameter estimates that maximizes prediction in that tuning cohort. In all methods, once the
166 SNPs and $\hat{b}_i$ have been decided, PLINK --*score* is used to calculate the PGS in the target
167 sample.

168

169 LD pruning and thresholding (P+T)[3]

170 In the P+T method GWAS summary statistics are pruned to be approximately independent
171 using a LD threshold, r^2. From this quasi-independent genome-wide SNP list, SNPs are

172 selected by thresholding on a pre-specified association p-value, P_t. We evaluated P+T as
173 implemented in Ricopili[31] which uses PLINK[32] to prune the SNP set using $r^2 = 0.1$ within
174 500 kb windows, and $P_t \in$ (5e-08, 1e-06, 1e-04, 1e-03, 0.01, 0.05, 0.1, 0.2, 0.5, 1), where $P_t = 1$
175 means that all SNPs from the LD-pruned list are included. In applications of P+T it is
176 common for results from the most associated P_t to be reported (including the application in
177 the software PRSice[33] which uses a continuous P_t range), but this approach utilises
178 information from the target cohort and hence introduces a form of winner's curse. Here, the
179 P_t threshold applied in target cohorts is the P_t threshold that maximised prediction in the
180 tuning cohort.
181

182 SBLUP[12]

183 SBLUP is a method that re-scales the GWAS SNP effect estimates using an external LD
184 reference panel to transform the ordinary least-squares estimates to approximate the best
185 linear unbiased prediction (BLUP) solutions. This method assumes an infinitesimal model
186 where SNP effects are drawn from a normal distribution. All genome-wide SNPs are used to
187 build the PGS. Hence, for example, consider a genomic region with a single causal variant
188 but with many SNPs in the region correlated with the causal variant and correlated with each
189 other. In this case the effect size estimate is "smeared" across the correlated SNPs, but with
190 the total contribution to risk expected to represent the best estimate of the signal from the
191 underlying causal variant. This method is implemented within the software package GCTA[34].
192

193 LDpred and LDpred-inf [13]

194 While P+T uses arbitrary LD and p-value thresholds for selection of SNPs, LDpred tries to
195 optimise this step in a Bayesian framework. The method uses the GWAS summary statistics
196 and LD information from the external LD reference sample to infer the posterior mean effect

size of each SNP, conditioning on the SNP effect estimates of other correlated SNPs. This method assumes a point-normal prior on the distribution of SNP effects such that only a fraction of SNPs with non-zero estimated effects are selected for inclusion in the PGS. The default parameter setting for the fractions of causal SNPs (π, but denoted p in the original paper) were used in the tuning cohort: $\pi \in \{1$ (i.e. all SNPs), 0.3, 0.1, 0.03, 0.01, 0.003, and 0.001$\}$, with an LD radius of M/3000 (M is the number of SNPs) to obtain local LD information, as suggested by the authors[13]. The π value that maximised the prediction in the tuning sample was applied in the target sample; the π value can differ between target cohorts even though the same tuning cohort is used, reflecting the properties of the discovery sample which may change with each left-out target sample. When π=1 the method is called LDpred-Inf and is equivalent to SBLUP (the concordance of results was checked, **Table S7**).

LDpred-funct[14]

LDpred-funct is an extension of the LDpred-inf (SBLUP equivalent) model but leverages trait-specific functional enrichments relative to the baseline-LD model[35] to up/down-weight SNP effects. The functional annotations include coding, conserved, regulatory and LD-related annotation. In the baseline-LD model, the enrichment of each category is jointly calculated via stratified LD score regression[36]. LDpred-funct has a non-infinitesimal model version, but in pilot analyses we found LDpred-Inf performed better than LDpred and hence only considered the LDpred-funct infinitesimal model. Thus, we continued only with the infinitesimal model version.

PRS-CS and PRS-CS-auto[15]

PRS-CS is also built under a Bayesian regression framework. Unlike LDpred which assumes a point-normal distribution as a prior, which is discrete, PRS-CS assumes a continuous

222 shrinkage prior on the SNP effects. PRS-CS was implemented using the default setting and
223 with the LD reference panel provided with the PRS-CS software, which is computed using
224 the 1000 Genomes samples and HapMap3 SNPs. In PRS-CS, for the global scaling parameter
225 which is applied to all SNP effects ϕ, the search grid is $\phi^{1/2} \in \{0.0001, 0.001, 0.01, 0.1, 1\}$,
226 The ϕ that produces the best predictive performance in a tuning data set is selected for use in
227 the target sample. In PRS-CS-auto, ϕ is automatically learnt from GWAS summary statistics
228 and no tunning sample is needed.
229

230 SBayesR

231 SBayesR is a method that re-scales the GWAS SNP effect estimates based on Bayesian
232 multiple regression. SBayesR assumes that the standardised SNP effects are drawn from a
233 mixture of four zero-mean normal distributions with different variances (one of the variances
234 is zero, with a probability of π_1), indicating that only a fraction of SNPs ($1-\pi_1$) have non-zero
235 estimated effects which contribute to the phenotype. Moreover, the contributions of SNPs in
236 different distributions differ because of different variances. Here, we evaluated SBayesR in
237 the default setting. For the LD reference, we used the same sparse LD matrix as the one used
238 in Lloyd-Jones et al.[16], where the LD matrix was built based on the HapMap3 SNPs of
239 randomly selected and unrelated 50K UK Biobank individuals. Whereas LDpred estimates
240 π from a tuning sample, SBayesR estimates π from the GWAS discovery sample, so no
241 tuning sample is needed.
242

243 Evaluation of out-of-sample prediction

244 The accuracy of prediction in each target cohort was quantified by 1) Area under the receiver
245 operator characteristic curve (AUC; R library pROC). AUC can be interpreted as a
246 probability that a case ranks higher than a control. 2) The proportion of variance on the

247 liability scale explained by PGS[37]. We used the population lifetime risk of SCZ and MDD as
248 1% and 15% respectively to convert the variance explained in a linear regression to the
249 liability scale[20; 23; 38]. 3) Odds ratio (OR) of tenth PGS decile relative to the first decile. 4)
250 Odds ratio of tenth PGS decile relative to those ranked in the middle of the PGS distribution,
251 which is calculated as the average of OR of tenth decile relative to fifth and sixth decile. 5)
252 Standard deviation unit increase in cases. The PGS in each target cohort were scaled by
253 standardising the PGS of controls and applying the standardisation to cases:
254 $\frac{PGS_{case}-mean(PGS_{control})}{SD(PGS_{control})}$, where SD is standard deviation. This does not impact PGS
255 evaluation statistics but simply means that PGS are in SD units for all cohorts. We compare
256 the median value for evaluation statistics 3 and 4, because they are significantly different
257 from a normal distribution based on a Shapiro-Wilk Normality Test. The regression analyses
258 for evaluation statistics 2-4 include 6 ancestry principal components as covariates. These
259 covariates are not included in the AUC model and the standard deviation unit increase in
260 cases model. To illustrate the impact on results, for SCZ given the SBayesR mean variance in
261 liability of 9.9% and lifetime risk of 0.01 the AUC expected from normal distribution
262 theory[39] is 0.730, compared to the mean reported of 0.733. For MDD given the variance in
263 liability of 4.0% and lifetime risk of 0.15 the expected AUC is 0.603 compared to the mean
264 reported of 0.601.
265

266 ## Results

267 Prediction evaluation statistics based on recent PGS methods applied to SCZ across 30 study
268 cohorts (**Figure 1**, **Table S3** and **S4**), and to MDD across 26 cohorts (**Figure 2, Table S5** and
269 **S6**), show higher values for all methods over the benchmark method, P+T. The evaluation
270 statistics include i) area under the receiver operator characteristic curve (AUC) which can be

271 interpreted as the probability that a case ranks higher than a control, when the case and
272 control are randomly drawn; ii) mean difference between cases and controls expressed in
273 PGS standard deviation (SD) units of controls, after standardization of the PGS so that
274 controls in each target cohort have a mean of 0 and a SD of 1; iii) variance in liability
275 explained by the PGS; iv) Odds ratio of the top 10% ranked on PGS relative to the bottom
276 10%; v) Odds ratio of the top 10% ranked on PGS relative to those ranking in the middle of
277 the PGS distribution; vi) difference between mean of PGS in the top 10% of cases and mean
278 PGS in top 10% of controls.
279
280 There is variability in prediction statistics across target cohorts which is not a reflection of
281 sample size (**Figure S1 and Table S4** for SCZ, **Figure S2 and Table S6** for MDD). To
282 provide a benchmark in terms of power, we note that for SCZ, the mean difference in PGS
283 between cases and controls for the P+T method is 0.73 standard deviation units of the control
284 sample (SDU). A sample size of only 42 cases and 42 controls has 95% power to detect this
285 difference at a nominal significance threshold of 0.05; all SCZ cohorts are bigger than this.
286 For MDD, the mean difference in PGS SDU between cases and controls for the P+T method
287 is 0.30, and the power calculation requires a sample size of 290 cases and 290 controls to
288 detect this difference; 20 (77%) of the MDD cohorts achieve this effective size. Hence, the
289 SCZ and MDD cohorts are well-powered for PGS evaluation.
290
291 The correlations of PGS between different methods are high (**Table S7**), but are lowest
292 between P+T and other methods (minimum 0.67). In contrast, the correlations between the
293 recent methods are always > 0.83. In theory, LDpred-Inf and SBLUP are the same method. In
294 practice, there are differences in implementation (e.g., different input parameters associated
295 with definition of LD window), generating a correlation 0.977. The differences in prediction

296 evaluation statistics between methods are small. For SCZ the AUC for all recent methods
297 other than PRS-CS-auto are significantly higher than the P+T method after Bonferroni
298 correction (p-value < 0.0018=0.05/28 (28 pair-wise comparisons between 8 methods), two-
299 tailed Student's t-test). For MDD none of the differences between methods were significant.
300 For both SCZ and MDD, regardless of tuning cohorts SBayesR showed relatively better
301 performance (on average across target cohorts) than other methods on all statistics, where
302 other recent methods performed similarly (**Figures 1 and 2**). For variance explained on the
303 liability scale, the P+T PGS explained a mean of 6.8% across cohorts for SCZ. For SBayesR,
304 the mean was 9.9% for variance explained in liability, an increase of 46%. For MDD
305 although the variance explained is lower in absolute terms, 2.8% for P+T vs 4.0% for
306 SBayesR; the latter represents a 43% increase.
307
308 We provide several evaluation statistics that focus on those in the top 10% of PGS, because
309 clinical utility of PGS for psychiatric disorders is likely to focus on individuals that are in the
310 top tail of the distribution of predicted genetic risk. The odds ratio for top vs bottom decile
311 are large, ranging from 13.8 for P+T to 22.5 for SBayesR for SCZ and 3 to 4 for MDD.
312 While these top vs bottom decile odds ratios (**Figure 1c** and **2c**) are much larger than the
313 odds ratio obtained by using PGS to screen a general population (**Figure 1d** and **2d**) or
314 patients in a healthcare system to identify people at high risk[40; 41], these comparisons are
315 useful for research purposes, which could for example make cost-effective experimental
316 designs focussing on individuals with high vs low PGS.[42] The odds ratio of top 10% vs
317 middle 10% are much less impressive, up to median of 5.5 for SCZ and 2 for MDD, but more
318 fairly represents the value of PGS in population settings. These values can be benchmarked
319 against risk in 1st degree relatives of those affected, which are of the order of 8 for SCZ and 2
320 for MDD; low values are always expected for MDD because it is more common (lifetime risk

321 ~15% compared to ~1% for SCZ). The odds ratio values are particularly high for some
322 cohorts (**Table S4**), because in some SCZ cohorts the bottom 10% include very few or no
323 cases, especially in cohorts with relatively small sample sizes. Since the PGS are normally
324 distributed, as expected the mean PGS for controls in the top 10% PGS is ~1.75 SD units
325 (K=0.10, t=qnorm(1-K), z=dnorm(t), mean value of top 10% of a normal distribution =z/K),
326 whereas the top 10% of cases have mean value of 2.63 control sample SD units for SCZ
327 cases and 2.09 control sample SD units for MDD cases, using SBayesR. These mean values
328 of the top 10% in cases equate to expectations from the population of the top 1.1% and top
329 4.7% for SCZ and MDD, respectively.
330
331 **The impact of tuning cohort.** Three methods (i.e., P+T, LDpred and PRC-CS) use tuning
332 cohorts to determine key parameters for application of the method into the target cohorts.
333 Tuning parameters impact results in two ways. First, the parameters may be dependent on the
334 choice of tuning cohort. Second, the discovery GWAS sample may be reduced in size (and
335 hence power) if a tuning cohort needs to be excluded from the discovery GWAS. In all our
336 analyses the tuning cohort is excluded from all GWAS discovery samples so that GWAS
337 discovery sample is not variable across methods for each target cohort. A sensitivity analyses
338 that used the SCZ cohorts of msaf (Ncase= 327, Ncontrol= 139), gras (Ncase= 1086,
339 Ncontrol= 1232) or swe6 (Ncase= 1094, Ncontrol= 1219) as the tuning sample instead of
340 cohort lie2 (Ncase= 137, Ncontrol= 269) show that the tuning cohort can have considerable
341 impact (**Figure 3 and Figures S3-5**). In our results, the tuning cohort that generates higher
342 PGS is method dependent and differs between cohorts. Although methods SBLUP, LDpred-
343 Inf, LDpred-funct, PRS-CS-auto and SBayesR require no tuning cohort, they serve as a
344 benchmark, since the differences in their results reflect differences in the changed discovery

345 samples (e.g., msaf is in the discovery sample, when lie2 is the tuning cohort, and *vice versa*),
346 as well as the stochasticity inherent in the Gibbs sampling of Bayesian methods.
347
348 **The impact of MAF/INFO threshold**. A MAF threshold of 0.1 and a INFO threshold of 0.9
349 are used to be consistent with applications in the PGC SCZ[23] and PGC MDD[20] studies, which
350 had been imposed recognising that these thresholds generated more robust PGS results than
351 using lower threshold values. In the second sensitivity analysis applied to the SCZ data, the
352 MAF threshold was relaxed to 0.05 (**Figure 4**) and to 0.01 (**Figure S6**). The prediction
353 evaluation statistics increase for some cohorts and decrease for others. SBLUP, PRS-CS,
354 PRS-CS-auto and SBayesR are less affected than P+T, LDpred-Inf, LDpred-funct and
355 LDpred. For QC threshold of MAF < 0.01, the differences in AUC have a similar trend
356 compared to using MAF<0.05, but with greater variability (**Figure S6**). The effects of MAF
357 thresholds vary between cohorts, although the use of lower MAF threshold tends to generate
358 higher AUC for the larger target samples. Across target cohorts, different evaluation statistics
359 were almost identical when including less common SNPs (**Table S3**). Relaxing the INFO
360 score to 0.3 has a negligible effect (**Figure S7**).

361 Discussion

362 Comparison of PGS risk prediction methods showed that all recent methods had higher
363 prediction evaluation statistics over the benchmark P+T method for SCZ and MDD. While
364 the differences between the recent methods were small, we found that SBayesR consistently
365 ranked highest. Given that the PGS is a sum of many small effects, a normal distribution of
366 PGS in a population is expected (and observed **Figures S8-S11**). In idealised data, such as
367 the relatively simple simulation scenarios usually considered in method development, all
368 evaluation statistics should rank the same, but with real data sets this is not guaranteed. This

369 is the motivation for considering a range of evaluation statistics. Our focus on statistics for
370 those in the top 10% of PGS is partly motivated by potential clinical utility. In the context of
371 psychiatry, it is likely that this will focus on people presenting in a prodromal state with
372 clinical symptoms that have not yet crystallised to a specific diagnosis[10; 43]. High PGS in
373 those presenting to clinics could help tilt the clinical decision-making towards closer
374 monitoring or earlier intervention. Since a genetic-based predictor only predicts part of the
375 risk of disease, and since a PGS only predicts part of the genetic contribution to disease it is
376 acknowledged that PGS cannot be fully accurate predictors. Nonetheless, PGS, in
377 combination with clinical risk factors, could make a useful contribution to risk prediction.
378
379 In sensitivity analyses that used different quality criteria for SNPs e.g. MAF of 0.01 vs 0.05,
380 INFO of 0.3 vs 0.9, we concluded that, currently, there is little to be gained in PGS from
381 including SNPs with MAF < 0.10 and INFO < 0.9 for the diseases/dataset studied (**Table S8**
382 and **S9**). This result may seem counter-intuitive since variants with low MAF are expected to
383 play an important role in common disease, and some may be expected to have larger effect
384 sizes than more common variants[44; 45]. However, sampling variance is a function of allele
385 frequency (var(y)/(2*MAF(1-MAF)*n), where n is sample size), such that a variant of MAF
386 =0.01 has sampling variance 25 times greater than a variant of MAF=0.5. Moreover, in real
387 data sets cohort sample size and technical artefacts can accumulate to increase error in effect
388 size estimates particularly of low frequency variants. Our conclusion that little is gained from
389 including variants of MAF < 0.1 and reducing INFO threshold needs to be revisited as larger
390 discovery samples and larger target cohorts accumulate.
391
392 For both SCZ and MDD, while all recent methods had similar performance, SBayesR saw the
393 highest prediction accuracy in most of the comparisons. Although SCZ and MDD both have a

394 highly polygenic genetic architecture, we have recently shown that SBayesR outperforms
395 other methods for two less polygenic diseases, Alzheimer's[46] (which includes the *APOE*
396 locus which has a very large effect size) and ALS[47] (for which there is evidence of greater
397 importance of low MAF variants compared to SCZ[48]). The original SBayesR publication
398 showed that in both simulations and applications to real data, the method performed well
399 across a range of traits with different underlying genetic architectures, which is because
400 SBayesR can fit essentially any underlying architecture and other methods are special cases
401 of the SBayesR model, except PRS-CS which uses different distributional approaches (**Table**
402 **1**). We note that we did not consider a version of P+T that has been shown to have higher out
403 of sample prediction compared to the standard implementation[11]. This method conducts a
404 grid search in a tuning cohort to determine LD r^2 and INFO score thresholds for SNPs as well
405 as the p-value threshold. We chose to implement only the basic, commonly used P+T
406 method, and specifically as implemented in published PGC studies. Moreover, many of the
407 methods implemented here address optimum SNP selection from a methodological approach
408 rather than grid search approach. We note that here we only considered the infinitesimal
409 model version of LDpred-funct, because we have already found no advantage of LDpred over
410 LDpred-inf in the preliminary analyses. For traits and diseases of other genetic architecture
411 parameters of LDpref-funct should be investigated, although in the updated LDpred-funct
412 preprint[49], SBayesR was found to perform well across a range of quantitative and binary
413 traits. We do note that SBayesR expects effect size estimates and their standard errors to have
414 properties consistent with the sample size and with the LD patterns imposed from an external
415 reference panel. If GWAS summary statistics have non-ideal properties (perhaps resulting
416 from meta-analysis errors or approximations) then SBayesR may not achieve converged
417 solutions. Last, we note that the comparison of methods uses only study samples of European
418 ancestry. More research is needed to understand the properties of prediction methods within

419 other ancestries and across ancestries, given potential differences in genetic architectures (in
420 terms of number, frequencies and effect sizes of causal variants) and LD between measured
421 variants and causal variants[50; 51]. Such research is dependent on the availability of large
422 GWAS data sets from non-European ancestries; currently there is considerable effort to
423 increase GWAS sample collection across world-wide population groups to address this
424 concern[50-52].
425
426 All recent methods are compared using their default parameters settings. An optimum setting
427 of each method could potentially increase the prediction accuracy. For example, in sensitivity
428 analyses we found that LDpred sees higher prediction accuracy when increasing the length of
429 MCMC chain, while PRS-CS-auto and SBayesR results are not impacted by increasing the
430 MCMC chain length beyond the default settings (**Table S10**). This result agrees with the
431 recent revision of LDpred, LDpred2[53]. The underlying model and assumptions about the SNP
432 effect distribution are unchanged, but higher prediction accuracy is reported for longer
433 MCMC chain and larger LD windows. Most likely the optimum parameter settings are trait
434 (genetic architecture) dependent[11]. Hence, we conclude that a key advantage of SBayesR is
435 that there is no need for the user to tune or select model or software parameters. Moreover, it
436 does not need a tuning cohort to derive SNP effect weights but learns the genetic architecture
437 from the properties of the GWAS results. A third key advantage of SBayesR is computational
438 speed. Using one CPU, it takes approximately 2 hours to generate SNP weights based on
439 each discovery sample and predict into the left-out-cohort, compared to PRS-CS which needs
440 40 hours using 5 CPUs (the CPU number is fixed in the PRS-CS software). Last, given that
441 SBayesR uses only HapMap3 SNPs that are mostly well-imputed it should be possible to
442 provide these SBayesR SNP weights as part of a GWAS pipeline to apply in external target
443 samples.

Supplemental Data

The Supplemental Data include 11 figures and consortium members.

Acknowledgements

We acknowledge funding from the National health and Medical Research Council (1173790,1078901,108788 (NRW),1113400 (NRW, PMV)) and the Australian Research Council (FL180100072 (PMV)).

This work would not have been possible without the contributions of the investigators who comprise the PGC-SCZ and PGC-MDD working group. For a full list of acknowledgments of all individual cohorts included in PGC-SCZ and PGC-MDD, please see the original publications. The PGC has received major funding from the US National Institute of Mental Health and the US National Institute of Drug Abuse (U01 MH109528 and U01 MH1095320).

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 AAHRPP-accredited institutional review board.

The Münster cohort was funded by the German Research Foundation (DFG, grant FOR2107 DA1151/5-1 and DA1151/5-2 to U.D.; SFB-TRR58, Projects C09 and Z02 to U.D.) and the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grant Dan3/012/17 to U.D.).

Some data used in this study were obtained from dbGaP. dbGaP accession phs000021: funding support for the Genome-Wide Association of Schizophrenia Study was provided by the National Institute of Mental Health (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

468 MH79469, and U01 MH79470), and the genotyping of samples was provided through the
469 Genetic Association Information Network (GAIN). Samples and associated phenotype data
470 for the Genome-Wide Association of Schizophrenia Study were provided by the Molecular
471 Genetics of Schizophrenia Collaboration (principal investigator P. V. Gejman, Evanston
472 Northwestern Healthcare (ENH) and Northwestern University, Evanston, IL, USA). dbGaP
473 accession phs000196: this work used in part data from the NINDS dbGaP database from the
474 CIDR: NGRC PARKINSON'S DISEASE STUDY. dbGaP accession phs000187: High-
475 Density SNP Association Analysis of Melanoma: Case–Control and Outcomes Investigation.
476 Research support to collect data and develop an application to support this project was
477 provided by P50 CA093459, P50 CA097007, R01 ES011740, and R01 CA133996 from the
478 NIH.
479 Statistical analyses were carried out on the Genetic Cluster Computer
480 (http://www.geneticcluster.org) hosted by SURFsara and financially supported by the
481 Netherlands Scientific Organization (NWO 480-05-003) along with a supplement from the
482 Dutch Brain Foundation and the VU University Amsterdam.

483 Declaration of Interests

484 The authors declare no competing interests.

485 References

486 1. The International Schizophrenia Consortium. (2009). Common polygenic variation
487 contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748-752.
488 2. Palk, A.C., Dalvie, S., De Vries, J., Martin, A.R., and Stein, D.J. (2019). Potential use of
489 clinical polygenic risk scores in psychiatry–ethical implications and communicating
490 high polygenic risk. Philos. Ethics Humanit. Med. 14, 4.
491 3. Wray, N.R., Goddard, M.E., and Visscher, P.M. (2007). Prediction of individual genetic risk
492 to disease from genome-wide association studies. Genome Res. 17, 1520-1528.
493 4. Jenkins, M.A., Win, A.K., Dowty, J.G., MacInnis, R.J., Makalic, E., Schmidt, D.F., Dite, G.S.,
494 Kapuscinski, M., Clendenning, M., and Rosty, C. (2019). Ability of known

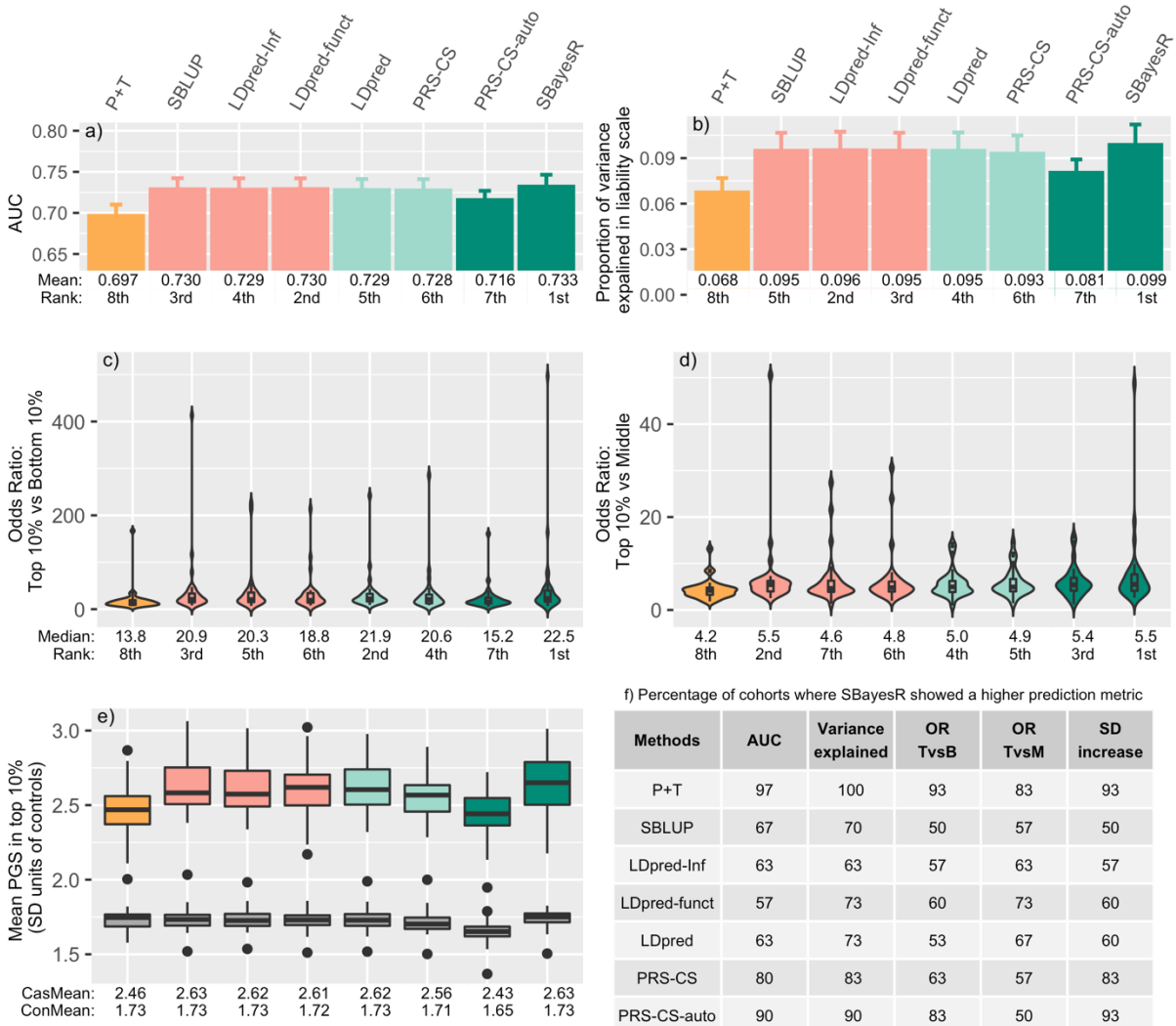
495 susceptibility SNPs to predict colorectal cancer risk for persons with and without a
496 family history. Fam. Cancer 18, 389-397.
497 5. Lee, A., Mavaddat, N., Wilcox, A.N., Cunningham, A.P., Carver, T., Hartley, S., de Villiers,
498 C.B., Izquierdo, A., Simard, J., and Schmidt, M.K. (2019). BOADICEA: a comprehensive
499 breast cancer risk prediction model incorporating genetic and non-genetic risk
500 factors. Genet. Med. 21, 1708.
501 6. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P.,
502 Lander, E.S., Lubitz, S.A., and Ellinor, P.T. (2018). Genome-wide polygenic scores for
503 common diseases identify individuals with risk equivalent to monogenic mutations.
504 Nat. Genet. 50, 1219-1224.
505 7. Lloyd-Jones, D.M., Wilson, P.W.F., Larson, M.G., Beiser, A., Leip, E.P., D'Agostino, R.B., and
506 Levy, D. (2004). Framingham risk score and prediction of lifetime risk for coronary
507 heart disease. Am. J. Cardiol. 94, 20-24.
508 8. McCarthy, M.I., and Mahajan, A. (2018). The value of genetic risk scores in precision
509 medicine for diabetes. Expert Rev. Precis. Med. Drug Dev. 3.
510 9. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of
511 polygenic risk scores. Nat. Rev. Genet. 19, 581.
512 10. Murray, G.K., Lin, T., Austin, J., McGrath, J.J., Hickie, Ian B., and Wray, N.R. (2020).
513 Polygenic risk scores - could they be useful in psychiatry? Submitted.
514 11. Privé, F., Vilhjálmsson, B.J., Aschard, H., and Blum, M.G.B. (2019). Making the most of
515 Clumping and Thresholding for polygenic scores. Am. J. Hum. Genet. 105, 1213-1221.
516 12. Robinson, M.R., Kleinman, A., Graff, M., Vinkhuyzen, A.A.E., Couper, D., Miller, M.B.,
517 Peyrot, W.J., Abdellaoui, A., Zietsch, B.P., and Nolte, I.M. (2017). Genetic evidence of
518 assortative mating in humans. Nat. Hum. Behav. 1, 0016.
519 13. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese,
520 G., Loh, P.-R., Bhatia, G., and Do, R. (2015). Modeling linkage disequilibrium
521 increases accuracy of polygenic risk scores. Am. J. Hum. Genet. 97, 576-592.
522 14. Marquez-Luna, C., Gazal, S., Loh, P.-R., Furlotte, N., Auton, A., Price, A.L., and andMe
523 Research, T. (2018). Modeling functional enrichment improves polygenic prediction
524 accuracy in UK Biobank and 23andMe data sets. bioRxiv, 375337.
525 15. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., and Smoller, J.W. (2019). Polygenic prediction via
526 Bayesian regression and continuous shrinkage priors. Nat. Commun. 10, 1776.
527 16. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H.,
528 Zheng, Z., Magi, R., and Esko, T. (2019). Improved polygenic prediction by Bayesian
529 multiple regression on summary statistics. bioRxiv.
530 17. Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic
531 risk prediction models for stratified disease prevention. Nat. Rev. Genet. 17, 392.
532 18. The International Schizophrenia Consortium. (2020). Manuscript in preparation.
533 19. Pardiñas, A.F., Holmans, P., Pocklington, A.J., Escott-Price, V., Ripke, S., Carrera, N.,
534 Legge, S.E., Bishop, S., Cameron, D., and Hamshere, M.L. (2018). Common
535 schizophrenia alleles are enriched in mutation-intolerant genes and in regions under
536 strong background selection. Nat. Genet. 50, 381-389.
537 20. Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A.,
538 Adams, M.J., Agerbo, E., Air, T.M., and Andlauer, T.M.F. (2018). Genome-wide
539 association analyses identify 44 risk variants and refine the genetic architecture of
540 major depression. Nat. Genet. 50, 668.

541 21. Howard, D.M., Adams, M.J., Clarke, T.-K., Hafferty, J.D., Gibson, J., Shirali, M., Coleman,
542 J.R.I., Hagenaars, S.P., Ward, J., and Wigmore, E.M. (2019). Genome-wide meta-
543 analysis of depression identifies 102 independent variants and highlights the
544 importance of the prefrontal brain regions. Nat. Neurosci. 22, 343.
545 22. Trzaskowski, M., Mehta, D., Peyrot, W.J., Hawkes, D., Davies, D., Howard, D.M., Kemper,
546 K.E., Sidorenko, J., Maier, R., and Ripke, S. (2019). Quantifying between-cohort and
547 between-sex genetic heterogeneity in major depressive disorder. American Journal
548 of Medical Genetics Part B: Neuropsychiatric Genetics 180, 439-447.
549 23. Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological
550 insights from 108 schizophrenia-associated genetic loci. Nature 511, 421-427.
551 24. Howard, D.M., Adams, M.J., Shirali, M., Clarke, T.-K., Marioni, R.E., Davies, G., Coleman,
552 J.R.I., Alloza, C., Shen, X., and Barbu, M.C. (2018). Genome-wide association study of
553 depression phenotypes in UK Biobank identifies variants in excitatory synaptic
554 pathways. Nat. Commun. 9, 1-10.
555 25. Hyde, C.L., Nagle, M.W., Tian, C., Chen, X., Paciga, S.A., Wendland, J.R., Tung, J.Y., Hinds,
556 D.A., Perlis, R.H., and Winslow, A.R. (2016). Identification of 15 genetic loci
557 associated with risk of major depression in individuals of European descent. Nat.
558 Genet. 48, 1031.
559 26. Banda, Y., Kvale, M.N., Hoffmann, T.J., Hesselson, S.E., Ranatunga, D., Tang, H., Sabatti,
560 C., Croen, L.A., Dispensa, B.P., and Henderson, M. (2015). Characterizing
561 race/ethnicity and genetic ancestry for 100,000 subjects in the Genetic Epidemiology
562 Research on Adult Health and Aging (GERA) cohort. Genetics 200, 1285-1295.
563 27. Pedersen, C.B., Bybjerg-Grauholm, J., Pedersen, M.G., Grove, J., Agerbo, E., Baekvad-
564 Hansen, M., Poulsen, J.B., Hansen, C.S., McGrath, J.J., and Als, T.D. (2018). The
565 iPSYCH2012 case–cohort sample: new directions for unravelling genetic and
566 environmental architectures of severe mental disorders. Mol. Psychiatry 23, 6-14.
567 28. Ripke, S., Wray, N.R., Lewis, C.M., Hamilton, S.P., Weissman, M.M., Breen, G., Byrne,
568 E.M., Blackwood, D.H.R., Boomsma, D.I., and Cichon, S. (2013). A mega-analysis of
569 genome-wide association studies for major depressive disorder. Mol. Psychiatry 18,
570 497-511.
571 29. Smith, B.H., Campbell, A., Linksted, P., Fitzpatrick, B., Jackson, C., Kerr, S.M., Deary, I.J.,
572 MacIntyre, D.J., Campbell, H., and McGilchrist, M. (2012). Cohort Profile: Generation
573 Scotland: Scottish Family Health Study (GS: SFHS). The study, its participants and
574 their potential for genetic research on health and illness. Int. J. Epidemiol. 42, 689-
575 700.
576 30. Fernandez-Pujals, A.M., Adams, M.J., Thomson, P., McKechanie, A.G., Blackwood, D.H.R.,
577 Smith, B.H., Dominiczak, A.F., Morris, A.D., Matthews, K., and Campbell, A. (2015).
578 Epidemiology and heritability of major depressive disorder, stratified by age of
579 onset, sex, and illness course in Generation Scotland: Scottish Family Health Study
580 (GS: SFHS). PLoS One 10, e0142197.
581 31. Lam, M., Awasthi, S., Watson, H.J., Goldstein, J., Panagiotaropoulou, G., Trubetskoy, V.,
582 Karlsson, R., Frei, O., Fan, C.-C., De Witte, W., et al. (2019). RICOPILI: Rapid
583 Imputation for COnsortias PIpeLIne. Bioinformatics.
584 32. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015).
585 Second-generation PLINK : rising to the challenge of larger and richer datasets.
586 GigaScience. 4.

587 33. Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: polygenic risk score software.
588 Bioinformatics 31, 1466-1468.
589 34. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: A tool for genome-
590 wide complex trait analysis. Am. J. Hum. Genet. 88, 76-82.
591 35. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., Schoech, A.,
592 Bulik-Sullivan, B., Neale, B.M., and Gusev, A. (2017). Linkage disequilibrium–
593 dependent architecture of human complex traits shows action of negative selection.
594 Nat. Genet. 49, 1421.
595 36. Finucane, H.K., Bulik-sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-r., Anttila, V.,
596 Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional
597 annotation using genome-wide association summary statistics. Nat. Genet. 47, 1228-
598 1235.
599 37. Lee, S.H., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2012). A better coefficient of
600 determination for genetic profile analysis. Genet. Epidemiol. 36, 214-224.
601 38. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing
602 heritability for disease from genome-wide association studies. Am. J. Hum. Genet.
603 88, 294-305.
604 39. Wray, N.R., Lin, T., Austin, J., McGrath, J.J., Hickie, Ian B., Murray, G.K., and Visscher,
605 P.M. (2020). Polygenic risk scores – from basic science to clinical application: a
606 primer. JAMA Psychiatry (In press).
607 40. Zheutlin, A.B., Dennis, J., Karlsson Linnér, R., Moscati, A., Restrepo, N., Straub, P.,
608 Ruderfer, D., Castro, V.M., Chen, C.-Y., and Ge, T. (2019). Penetrance and pleiotropy
609 of polygenic risk scores for schizophrenia in 106,160 patients across four health care
610 systems. Am. J. Psychiatry 176, 846-855.
611 41. Binder, E.B. (2019). Polygenic Risk Scores in Schizophrenia: Ready for the Real World?
612 Am. J. Psychiatry 170, 783-784.
613 42. Rehbach, K., Zhang, H., Das, D., Abdollahi, S.S., Prorok, T., Ghosh, S., Weintraub, S.,
614 Genovese, G., Powell, S., and Lund, A. (2020). Publicly available hiPSC lines with
615 extreme polygenic risk scores for modeling schizophrenia. bioRxiv.
616 43. Perkins, D.O., Olde Loohuis, L., Barbee, J., Ford, J., Jeffries, C.D., Addington, J., Bearden,
617 C.E., Cadenhead, K.S., Cannon, T.D., and Cornblatt, B.A. (2020). Polygenic risk score
618 contribution to psychosis prediction in a target population of persons at clinical high
619 risk. Am. J. Psychiatry 177, 155-163.
620 44. Park, J.-H., Gail, M.H., Weinberg, C.R., Carroll, R.J., Chung, C.C., Wang, Z., Chanock, S.J.,
621 Fraumeni, J.F., and Chatterjee, N. (2011). Distribution of allele frequencies and effect
622 sizes and their interrelationships for common genetic susceptibility variants.
623 Proceedings of the National Academy of Sciences 108, 18026-18031.
624 45. Bomba, L., Walter, K., and Soranzo, N. (2017). The impact of rare and low-frequency
625 genetic variants in common disease. Genome Biol. 18, 77.
626 46. Zhang, Q., Visscher, P.M., and McRae, A.F. (2020). Risk prediction of late-onset
627 Alzheimer's disease implies an oligogenic architecture. Submitted.
628 47. Restuadi, R., Garton, F.C., Benyamin, B., and Lin, T. (2020). Polygenic Risk Score Analysis
629 for Amyotrophic Lateral Sclerosis leveraging Cognitive Performance, Educational
630 Attainment and Schizophrenia. Eur. J of Hum. Genet.
631 48. van Rheenen, W., Shatunov, A., Dekker, A.M., McLaughlin, R.L., Diekstra, F.P., Pulit, S.L.,
632 van der Spek, R.A., Võsa, U., de Jong, S., and Robinson, M.R. (2016). PARALS Registry.
633 SLALOM Group. SLAP Registry. FALS Sequencing Consortium. SLAGEN Consortium.

634 NNIPPS Study Group Genome-wide association analyses identify new risk variants
635 and the genetic architecture of amyotrophic lateral sclerosis. Nat. Genet. 48, 1043-
636 1048.
637 49. Márquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S.S., and Furlotte, N. LDpred-funct:
638 incorporating functional priors improves polygenic prediction accuracy in UK Biobank
639 and 23andMe data sets.
640 50. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019).
641 Clinical use of current polygenic risk scores may exacerbate health disparities. Nat.
642 Genet. 51, 584.
643 51. Peterson, R.E., Kuchenbaecker, K., Walters, R.K., Chen, C.Y., Popejoy, A.B., Periyasamy,
644 S., Lam, M., Iyegbe, C., Strawbridge, R.J., Brick, L., et al. (2019). Genome-wide
645 Association Studies in Ancestrally Diverse Populations: Opportunities, Methods,
646 Pitfalls, and Recommendations. Cell 179, 589-603.
647 52. Karczewski, K.J., and Martin, A.R. (2020). Analytic and Translational Genetics. Annu. Rev.
648 Biomed. Data Sci. 3.
649 53. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2020). LDpred2: better, faster, stronger.
650 BioRxiv.
651

Figure



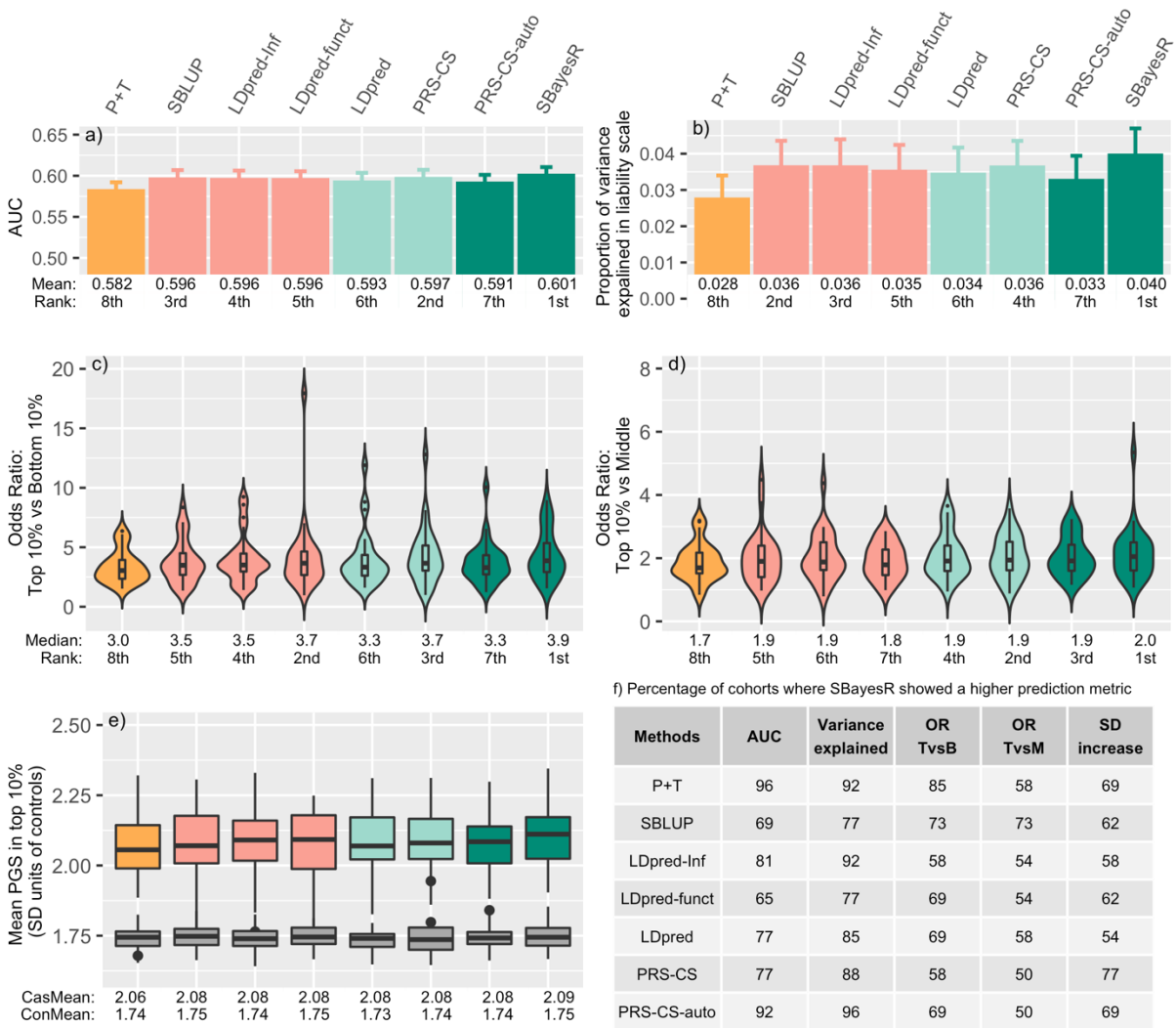
f) Percentage of cohorts where SBayesR showed a higher prediction metric

Methods	AUC	Variance explained	OR TvsB	OR TvsM	SD increase
P+T	97	100	93	83	93
SBLUP	67	70	50	57	50
LDpred-Inf	63	63	57	63	57
LDpred-funct	57	73	60	73	60
LDpred	63	73	53	67	60
PRS-CS	80	83	63	57	83
PRS-CS-auto	90	90	83	50	93

Figure 1. Results from prediction of SCZ case/control status using different PGS methods.

a) The area under curve (AUC) statistic. The AUC is a measure for the prediction accuracy, which indicates the probability that a case ranks higher than a control. The predictors were constructed from GWAS summary statistics of PGC-SCZ2+ excluding the target cohort and the tuning cohort (cohort 'lie2'). Each bar reflects the mean AUC across 30 target cohorts, the whiskers show the 95% confidence interval for comparing means. The number below each bar is the mean AUC estimated by each method, followed by its rank. P+T is the benchmark method which is shown in orange. Pink shows the methods that use

663 an infinitesimal model assumption. Light green shows the methods using a tuning cohort
664 to determine the genetic architecture of a trait. Dark green shows the methods learning the
665 genetic architecture from discovery sample, without using a tuning cohort.
666 b) The proportion of variance explained by PGS on the scale of liability, assuming a
667 population lifetime risk of 1%.
668 c) The odds ratio when considering the odds of being a case comparing the top 10% vs
669 bottom 10% of PGS. The number below each violin is the median OR estimated by each
670 method, followed by its rank.
671 d) The odds ratio when considering the odds of being a case comparing the top 10% vs those
672 in the middle of the PGS distribution, calculated as the averaged odds ratio of the top
673 10% ranked on PGS relative to the 5th decile and 6th decile. The number below each
674 violin is the median OR estimated by each method, followed by its rank.
675 e) The mean of the PGS for the top 10% cases (coloured boxes) and for the top 10% of
676 controls (grey boxes) in PGS standard deviation (SD) unit scale so that controls have
677 mean PGS of zero and SD of 1.
678 f) This table shows the percentages of the number of cohorts (out of 30) where SBayesR
679 showed a higher prediction metric compared to different methods. AUC: Area under
680 curve; Variance explained: The proportion of variance explained by PGS in liability
681 scale; OR TvsB: odds ratio, comparing the top 10% vs bottom 10% of PGS; OR TvsM:
682 odds ratio, comparing the top 10% vs those in the middle of the PGS distribution; SD
683 increase: standard deviation units increasing of the PGS for the top 10% cases.



f) Percentage of cohorts where SBayesR showed a higher prediction metric

Methods	AUC	Variance explained	OR TvsB	OR TvsM	SD increase
P+T	96	92	85	58	69
SBLUP	69	77	73	73	62
LDpred-Inf	81	92	58	54	58
LDpred-funct	65	77	69	54	62
LDpred	77	85	69	58	54
PRS-CS	77	88	58	50	77
PRS-CS-auto	92	96	69	50	69

Figure 2. Results from prediction of MDD case/control status using different PGS methods.

Similar to the caption of Figure 1, but the predictors were constructed from GWAS summary statistics of UK Biobank[15], 23andMe[16], GERA[17], iPSYCH[18], deCODE[19], GenScotland[20,21], PGC-MDD29 excluding the target cohort. The target cohorts comprised 26 of the 29 cohorts in MDD29. A cohort from Muenster, not included in the MDD29 was used as the tuning sample. The assumed population lifetime risk was 15% when estimating the proportion of variance explained by MDD PGS in liability scale.

Figure 3. Sensitivity analyses using different tuning cohorts.

Differences in AUC of a PGS 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 cohort (i.e. msaf, sew6 or gras), compared to 'lie2'. The MAF QC threshold is 0.1. Note: SBLUP, LDpred-Inf and LDpred-funct, PRS-CS-auto and SBayesR do not need a tuning cohort, but serve as a benchmark to methods which need a tuning cohort. These methods differ when a different tuning cohort is left out, because the discovery GWAS also changes.

Figure 4. Sensitivity analyses using different MAF quality control thresholds. Differences in AUC of a PGS method when using different MAF QC 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 MAF<0.05 and MAF <0.1 as a QC threshold. The tuning cohort is 'lie2'.

Table

711 Table 1. Summary of methods used to generate PGS

Method	Distribution of SNP effects	Tuning sample	Pre-defined parameters	Parameters estimated in tuning sample
P+T	None	Yes	-	P value threshold
SBLUP	$\beta \sim N(0, \frac{h_g^2}{m})$ h_g^2: SNP-heritability, m: number of SNPs	No	M, h_g^2, LD radius in kb	-
LDpred-Inf	Same as SBLUP	No	Sample size, LD radius in number of SNPs	-
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.	No	h_g^2, Per-SNP heritability estimated from stratified LDSC, Sample size, LD radius in number of SNPs	-
LDpred	$\beta_j \sim \begin{cases} N(0, \frac{h_g^2}{\pi m}), & \text{with probability of } \pi \\ 0, & \text{with probability of } 1-\pi \end{cases}$	Yes	Sample size, LD radius in number of SNPs	π
PRS-CS	$\beta_j \sim N(0, \frac{\sigma^2}{n}\psi_j)$ $\psi_j \sim G(a, \delta_j)$ $\delta_j \sim G(b, \phi)$ ϕ is a global scaling parameter. n is sample size G is a Gamma distribution	Yes	a, b, Sample size	ϕ
PRS-CS-auto	Same as PRS-CS,but estimates ϕ from the discovery GWAS.	No	a, b, Sample size	-

SBayesR	$\beta_j \mid \pi, \sigma_\beta^2 \sim \begin{cases} 0, & \text{with probability of } \pi_1 \\ N(0, \gamma_2 \sigma_\beta^2), & \text{with probability of } \pi_2 \\ & \vdots \\ N(0, \gamma_c \sigma_\beta^2), & \text{with probability of } 1 - \sum_{c=1}^{C-1} \pi_c \end{cases}$ $\sigma_\beta^2 \sim Inv - \chi^2(4)$ $\pi_i \sim Dir(\mathbf{1})$ γ_i are scaling parameters	No	γ_i	-

Distributions: *N:* normal, *G:* gamma, $Inv - \chi^2$: inverse chi-squared distribution, Dir: Dirichlet