Published in final edited form as:

Biol Psychiatry. 2021 November 01; 90(9): 611–620. doi:10.1016/j.biopsych.2021.04.018.

Correspondence: Naomi R Wray, Institute for Molecular Bioscience, The University of Queensland, Brisbane Qld 4072 Australia, T +61 7 3346 6374, naomi.wray@uq.edu.au.

The authors declare no competing interests.

The Schizophrenia Working Group of the Psychiatric Genomics Consortium is a collaborative co-author for this article. The individual authors are (affiliations are listed in the Supplement file) 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 & Michael C. O'Donovan

The Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium is a collaborative co-author for this article. The individual authors are (affiliations are listed in the Supplement file) 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

A comparison of ten polygenic score methods for psychiatric disorders applied across multiple cohorts

Guiyan Ni[1], **Jian Zeng**[1], **Joana A Revez**[1], **Ying Wang**[1], **Zhili Zheng**[1], **Tian Ge**[2], **Restuadi Restuadi**[1], **Jacqueline Kiewa**[1], **Dale R Nyholt**[3], **Jonathan R I Coleman**[4], **Jordan W Smoller**[2,5,6], **Schizophrenia Working Group of the Psychiatric Genomics Consortium**[7], **Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium**[8], **Jian Yang**[1,9], **Peter M Visscher**[1], **Naomi R Wray**[1,10]

[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]School of Life Sciences, Westlake University, Hangzhou, Zhejiang, 310024, China

[10]Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, 4072, Australia

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 (GWASs). PGS methods differ in which DNA variants are included and the weights assigned

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, EJC 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 AF 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, Nancy L Pedersen

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 (SCZ) and major depressive disorder (MDD) bring together many independently collected case-control cohorts. We used these resources (31K SCZ cases, 41K controls; 248K MDD cases, 563K 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 nine methods that model genetic architecture more formally: SBLUP, LDpred2-Inf, LDpred-funct, LDpred2, Lassosum, PRS-CS, PRS-CS-auto, SBayesR, MegaPRS) are compared.

Results: Compared to PC+T, the other nine methods give higher prediction statistics, MegaPRS, LDPred2 and SBayesR significantly so, up to 9.2% variance in liability for SCZ across 30 target cohorts, an increase of 44%. For MDD 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 comparison and are recommended in applications to psychiatric disorders.

Keywords

Polygenic scores; psychiatric disorders; schizophrenia; major depressive disorder; SBayesR; risk prediction; MegaPRS; LDpred2; PRS-CS; Lassosum

Introduction

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 (GWAS)(3). PGS can be calculated for any trait or disease with sufficiently powered GWAS (‘discovery samples’), and accuracy of PGS applied in independent GWAS ‘target samples’ will increase as discovery sample size increases. Since genetic factors only capture the genetic contribution to risk and since PGS only capture part of the genetic risk, PGS 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 PGS 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. Here, for simplicity, we assume 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 GWAS, 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 optimum method to decide which SNPs to select and what weights to allocate them, may differ between 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 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, not encountered in idealised 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 artefacts in genotype generation.

Here, we compare ten 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) and SBayesR(19), 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 which can be independently collected from the GWAS data, but from samples of matched ancestry.

Briefly, PC+T (P-value based clumping and thresholding, also known as the P+T or C+T method) 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. The 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). LDpred-funct and MegaPRS include functional annotation to SNPs to up/down weight 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 maximises prediction in the tuning cohort. MegaPRS utilises the BLD-LDAK model(22) where 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. Since LDpred2-auto is shown to perform similarly to LDpred2, we do not include it in comparisons made here. We apply these methods to data from the Psychiatric Genomics Consortium (PGC) working groups for schizophrenia (SCZ) (23, 24) and major depressive disorder (MDD)(12, 25, 26) (Tables S1 and S2). We select 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 psychiatric disorders which have all been shown to be highly polygenic (27). 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 PGS applied across multiple left-out target cohorts.

Materials and Methods

Data

All samples were of European ancestry with full details in the Supplementary Note, Table S1 and S2. Briefly, GWAS summary statistics were available from PGC SCZ for 37 European ancestry cohorts (24) (31K SCZ cases and 41K controls) of which 34 had individual level data available. PGS 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 (24) i.e., the target sample was excluded from the discovery sample as well as a sample selected to be a tuning sample. Analyses were repeated using four different tuning samples, two of which were large (swe6:2313; gras: 2318) and two were small (lie2:406; msaf:466). Similarly, GWAS MDD summary statistics were available from 248K cases and 563K controls(25), which included data from the 26 cohorts from PGC MDD with individual level data (15K cases and 24K 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(25), not included in the discovery GWAS was used as the tuning sample (N=1,679).

Baseline SNP selection

For baseline analyses, only SNPs with minor allele frequency (MAF) > 0.1 and imputation 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, except the method PC+T, which was conducted based on all imputed SNPs (8M in SCZ, and 13M in MDD).

Prediction methods

We define a PGS of an individual, j, as a weighted sum of SNP allele counts: $\sum_{i=1}^{m} \hat{b}_i x_{ij}$, 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 ten risk prediction methods, described in the Supplemental Note 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 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 are used.

Evaluation of out-of-sample prediction

The accuracy of prediction in each target cohort was quantified by 1) Area under the receiver operator characteristic curve (AUC; R library pROC(35)). 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(36). 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(25, 28, 37). 3) Odds ratio (OR) of tenth PGS decile relative to the first decile. 4) Odds ratio 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 PGS in each target cohort were scaled by standardising the PGS of controls and applying the standardisation 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 PGS are in SD units for all cohorts. The regression analyses for evaluation statistics 2–4 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 Supplementary Note).

Results

Prediction evaluation statistics based on all ten PGS methods and applied to SCZ across 30 study cohorts (Figure 1, Figure S1, Table S3 and S4), and to MDD across 26 cohorts (Figure S2, Table S5 and S6) are presented. There is variability in prediction statistics across target cohorts (as observed before(12, 28)) which is not a reflection of sample size (Figure S3 and Table S4 for SCZ, Figure S4 and Table S6 for MDD). Some significant associations were found from regression of prediction statistics on principal components (PCs) estimated from genome-wide SNPs (for SCZ Figure S3, but not MDD Figure S4), where the PCs capture both within-European ancestry and array differences between cohorts. The correlations of PGS between different methods are high (Table S7), but are lowest between PC+T and other methods (minimum 0.68). In contrast, the correlations between the other nine methods are always > 0.82. In theory, LDpred2-Inf and SBLUP are the same method. In practice, there are differences in implementation (e.g., different input parameters associated with definition of LD window) and although the correlation between their PGS is 0.974 the prediction accuracy is consistently higher for LDpred2-Inf. For SCZ, the AUC for all nine methods that directly model genetic architecture, other than PRS-CS-auto, are significantly higher than the PC+T method at the nominal level (Figure 1A). PGS from LDpred2, SBayesR and MegaPRS are significantly higher than the PC+T method after Bonferroni correction (p-value < 0.0011=0.05/45 (45 pairwise comparisons between 10 methods), one-tailed Student’s t-test). For MDD none of the differences between methods were significant (Figure S2A). For both SCZ and MDD across all statistics, regardless of tuning cohorts, LDpred2, SBayesR and MegaPRS, show relatively better performance (median across target cohorts) than other methods, although there is no significant difference between the nine 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 four tuning cohorts (Figure 1B), while it 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 *vs* 3.4% for MegaPRS, 3.5% for LDpred2 and 3.5% for SBayesR; the latter represents a 59% increase (Figure S2B).

We provide several evaluation statistics that focus on those in the top 10% of PGS, because clinical utility of PGS for psychiatric disorders is likely to focus on individuals that are in the top tail of the distribution of predicted genetic risk. The odds ratio for top *vs* bottom decile are 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 *vs* bottom decile odds ratios (Figure 1C and S2C) are much larger than the odds ratio obtained by using PGS to screen a general population (Figure 1D and 2D) or patients in a healthcare system to identify people at high risk(38, 39), these comparisons are useful for research purposes, which could, for example, make cost-effective experimental designs focussing on individuals with high *vs* low PGS(40). The odds ratio of top 10% *vs* middle 10% are much less impressive, up to median of 6 for SCZ and 2 for MDD, but more fairly represents the value of PGS in population settings. These values can be benchmarked against risk in 1st degree relatives of those affected, which are of the order of 8 for SCZ and 2 for MDD; low values are always expected for MDD because it is more common (lifetime risk ~15% compared to ~1% for SCZ). The odds ratio values are particularly high for some cohorts (Table S4), because in some SCZ cohorts the bottom 10% include very few or no cases, especially in cohorts with relatively small sample sizes.

The 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 into 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 power) if a tuning cohort needs to be excluded from the discovery GWAS. In all our analyses the tuning cohort is excluded from all GWAS discovery samples so that GWAS discovery sample is not variable across methods for each target cohort. Our results show that the tuning cohort can have considerable impact (Figures 1, 2). In our results, the tuning cohort that generates higher PGS is method dependent and differs between cohorts. For the methods that use tuning samples, the larger tuning samples (swe6 and gras) mostly generate higher prediction statistics compared to the two smaller tuning samples (lie and msaf), but the differences are not statistically significant. Although methods SBLUP, LDpred2-Inf, LDpred-funct, PRS-CS-auto and SBayesR require no tuning cohort, they serve as a benchmark, since 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.

The impact of MAF/INFO threshold.

A MAF threshold of 0.1 and a INFO threshold of 0.9 are used to be consistent with applications in the PGC SCZ(28) and PGC MDD(25) studies, which had been imposed recognising 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 increase for some cohorts and decrease for others (trends with sample size were not significant). PC+T is more impacted that the other nine methods. Across target cohorts, different evaluation statistics were almost identical when including less common SNPs (Table S3). Relaxing the INFO score to 0.3 has a negligible effect (Figure S5).

Discussion

Comparison of PGS risk prediction methods showed that all nine 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 nine 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 PGS in a population is expected (and observed Figures S6–9). In idealised data, such as the relatively simple simulation scenarios usually considered in method development, all evaluation statistics should rank the same, 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 PGS 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 specific to a diagnosis(11, 41). High PGS in those presenting to clinics could help contribute to clinical decision-making identifying individuals for closer monitoring or earlier intervention. Since a genetic-based predictor only predicts part of the risk of disease, and since a PGS only predicts part of the genetic contribution to disease it is acknowledged that PGS cannot be fully accurate predictors. Hence, the discriminative ability of PGS is low in the general population and the use of PGS in clinical settings requires evaluation including related ethical issues (42). Nonetheless, PGS, in combination with clinical risk factors, could make a useful contribution to risk prediction(41, 43, 44).

In sensitivity analyses that used different quality criteria for SNPs, e.g. MAF of 0.01 *vs* 0.05, INFO of 0.3 *vs* 0.9, we concluded that, currently, there is little to be gained in PGS from including SNPs with MAF < 0.10 and INFO < 0.9 for the diseases/dataset studied (Table S8 and S9). This result may seem counter-intuitive since variants with low MAF are expected to play an important role in common disease, and some may be expected to have larger effect sizes than more common variants(45, 46). 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 mean that technical artefacts can accumulate to increase error in effect size estimates particularly of low frequency variants. Our conclusion that little is gained from including variants of MAF < 0.1 and reducing INFO threshold needs to be revisited as larger individual cohorts in discovery samples and larger target cohorts accumulate. 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(47, 48).

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 to 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. Since the optimum LD threshold is likely to vary across genomic regions, the grid search approach is less appealing than the methods which 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 optimising this (Table S10). LDpred2 has a version that does not require a tuning sample, LDpred2-auto, but the authors showed the two methods give similar results. SBayesR assumes that the SNP effects are drawn from a mixture of four 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 term 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 the 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 (49) (which includes the *APOE* locus which has a very large effect size) and amyotrophic lateral sclerosis (50) (for which there is evidence of greater importance of low MAF variants compared to SCZ(51)). 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 four 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 maximises prediction in the tuning sample and then selects the best method amongst these. Hence, MegaPRS is a collection of the other methods and the SNP distribution selected varies depending on both tuning and target (Table S11). It selects BayesR 87% of the time when the tuning samples were large (otherwise BOLT-LMM) and selects Lasso 78% of the time when tuning samples were small. We implemented MegaPRS using the BLD-LDAK model recommended by the authors which assumes that the distribution of SNP effects depend 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 to LDpred2 and SBayesR that did not use functional annotations. Surprisingly, LDpred-funct method performed consistently less well than LDpred2-Inf, but this should be revisited as currently LDpred-funct is only available 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 (52). Consistent with our results, SBayesR attained a high prediction accuracy for MDD although they reported performance of SBayesR varied across traits. Since 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 non-ideal properties (perhaps resulting from meta-analysis errors or approximations) then SBayesR may not achieve converged solutions. SBayesR, in general, is more sensitive to any inconsistent properties between GWAS and LD reference samples than those methods that select hyperparameters based on cross-validation in a tuning sample, such as LDpred2 (15). We note that the LDpred-funct preprint reported SBayesR to perform 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 learns the genetic architecture from the properties of the GWAS results. Computationally it is also very efficient, using one CPU, it takes approximately 2 hours to generate SNP weights based on each discovery sample and predict into the left-out-cohort using a MCMC chain of 10,000 iterations (the computing time can be reduced by running a shorter chain since a negligible change in prediction accuracy was found after 4,000 iterations), which compares to PRS-CS: 40 hours using 5 CPUs, LDpred2: 5 hours using 15 CPUs, MegaPRS: 1 hours 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 parameters settings. An optimum setting of each method could potentially increase the prediction accuracy. Most likely the optimum parameter settings are trait (genetic architecture) dependent(13). Here, we find that all methods that more formally model the genetic architecture than PC+T perform better than the 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements and Disclosures

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 MH79469, and U01 MH79470), and the genotyping of samples was provided through the Genetic Association Information Network (GAIN). 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 (ENH) and Northwestern University, Evanston, IL, USA). dbGaP accession phs000196: this work used in part data from the NINDS dbGaP database from the CIDR: NGRC 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 P50 CA093459, P50 CA097007, R01 ES011740, and R01 CA133996 from the NIH.

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 (NWO 480-05-003) along with a supplement from the Dutch Brain Foundation and the VU University Amsterdam.

References

1. The International Schizophrenia Consortium (2009): Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 460:748–752. [PubMed: 19571811]
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. [PubMed: 30813945]
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. [PubMed: 17785532]
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. [PubMed: 32997097]
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. [PubMed: 31209717]
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. Genetics in medicine: official journal of the American College of Medical Genetics. 21:1708. [PubMed: 30643217]
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. [PubMed: 30104762]
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. The American journal of cardiology. 94:20–24. [PubMed: 15219502]
9. McCarthy MI, Mahajan A (2018): The value of genetic risk scores in precision medicine for diabetes. Taylor & Francis.
10. Torkamani A, Wineinger NE, Topol EJ (2018): The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 19:581. [PubMed: 29789686]
11. Murray GK, Lin T, Austin J, McGrath JJ, Hickie IB, Wray NR (2020): Could polygenic risk scores be useful in psychiatry?: A review. JAMA psychiatry.
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. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 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. [PubMed: 31761295]
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. BioRxiv.
16. Márquez-Luna C, Gazal S, Loh P-R, Kim SS, Furlotte N (2020): Ldpred-funct: Incorporating functional priors improves polygenic prediction accuracy in uk biobank and 23andme data sets. bioRxiv.
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. [PubMed: 28480976]

18. Ge T, Chen C-Y, Ni Y, Feng Y-CA, Smoller JW (2019): Polygenic prediction via bayesian regression and continuous shrinkage priors. Nat Commun. 10:1776. [PubMed: 30992449]
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. [PubMed: 30602773]
20. Zhang Q, Prive F, Vilhjalmsson BJ, Speed D (2020): Improved genetic prediction of complex traits from individual-level data or summary statistics. bioRxiv.
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. [PubMed: 27140283]
22. Speed D, Balding DJ (2019): Sumher better estimates the snp heritability of complex traits from summary statistics. Nat Genet. 51:277–284. [PubMed: 30510236]
23. The International Schizophrenia Consortium (2020): Manuscript in preparation.
24. 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. [PubMed: 29483656]
25. 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. [PubMed: 29700475]
26. 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. [PubMed: 30718901]
27. Sullivan PF, Geschwind DH (2019): Defining the genetic, genomic, cellular, and diagnostic architectures of psychiatric disorders. Cell. 177:162–183. [PubMed: 30901538]
28. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014): Biological insights from 108 schizophrenia-associated genetic loci. Nature. 511:421–427. [PubMed: 25056061]
29. Hyde CL, Nagle MW, Tian C, Chen X, Paciga SA, Wendland JR, et al. (2016): Identification of 15 genetic loci associated with risk of major depression in individuals of european descent. Nat Genet. 48:1031. [PubMed: 27479909]
30. Banda Y, Kvale MN, Hoffmann TJ, Hesselson SE, Ranatunga D, Tang H, et al. (2015): Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the genetic epidemiology research on adult health and aging (gera) cohort. Genetics. 200:1285–1295. [PubMed: 26092716]
31. Pedersen CB, Bybjerg-Grauholm J, Pedersen MG, Grove J, Agerbo E, Baekvad-Hansen M, et al. (2018): The ipsych2012 case–cohort sample: New directions for unravelling genetic and environmental architectures of severe mental disorders. Mol Psychiatry. 23:6–14. [PubMed: 28924187]
32. Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, Breen G, et al. (2013): A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatry. 18:497–511. [PubMed: 22472876]
33. Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, Kerr SM, et al. (2012): Cohort profile: Generation scotland: Scottish family health study (gs: Sfhs). The study, its participants and their potential for genetic research on health and illness. Int J Epidemiol. 42:689–700. [PubMed: 22786799]
34. Fernandez-Pujals AM, Adams MJ, Thomson P, McKechanie AG, Blackwood DHR, Smith BH, et al. (2015): Epidemiology and heritability of major depressive disorder, stratified by age of onset, sex, and illness course in generation scotland: Scottish family health study (gs: Sfhs). PLoS One. 10:e0142197. [PubMed: 26571028]
35. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. (2011): Proc: An open-source package for r and s+ to analyze and compare roc curves. BMC Bioinformatics. 12:1–8. [PubMed: 21199577]
36. Lee SH, Goddard ME, Wray NR, Visscher PM (2012): A better coefficient of determination for genetic profile analysis. Genet Epidemiol. 36:214–224. [PubMed: 22714935]
37. 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. [PubMed: 21376301]

38. 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. [PubMed: 31416338]

39. Binder EB (2019): Polygenic risk scores in schizophrenia: Ready for the real world? : Am Psychiatric Assoc.

40. 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.

41. 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. [PubMed: 31711302]

42. Wray NR, Lin T, Austin J, McGrath JJ, Hickie IB, Murray GK, et al. (2020): From basic science to clinical application of polygenic risk scores: A primer. JAMA psychiatry.

43. 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. [PubMed: 30309464]

44. 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. [PubMed: 27363508]

45. Park J-H, 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. Proceedings of the National Academy of Sciences. 108:18026–18031.

46. Bomba L, Walter K, Soranzo N (2017): The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 18:77. [PubMed: 28449691]

47. 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. [PubMed: 30926966]

48. 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. [PubMed: 31607513]

49. 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. 11:1–11. [PubMed: 31911652]

50. Restuadi R, Garton FC, Benyamin B, Lin T (2020): Polygenic risk score analysis for amyotrophic lateral sclerosis leveraging cognitive performance, educational attainment and schizophrenia. European Journal of Human Genetics. In press.

51. van Rheenen W, Shatunov A, Dekker AM, McLaughlin RL, Diekstra FP, Pulit SL, et al. (2016): Parals registry. Slalom group. Slap registry. Fals sequencing consortium. Slagen consortium. Nnipps study group genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat Genet. 48:1043–1048. [PubMed: 27455348]

52. Pain O, Glanville KP, Hagenaars SP, Selzam SP, Fürtjes AE, Gaspar HA, et al. (2020): Evaluation of polygenic prediction methodology within a reference-standardized framework. bioRxiv.

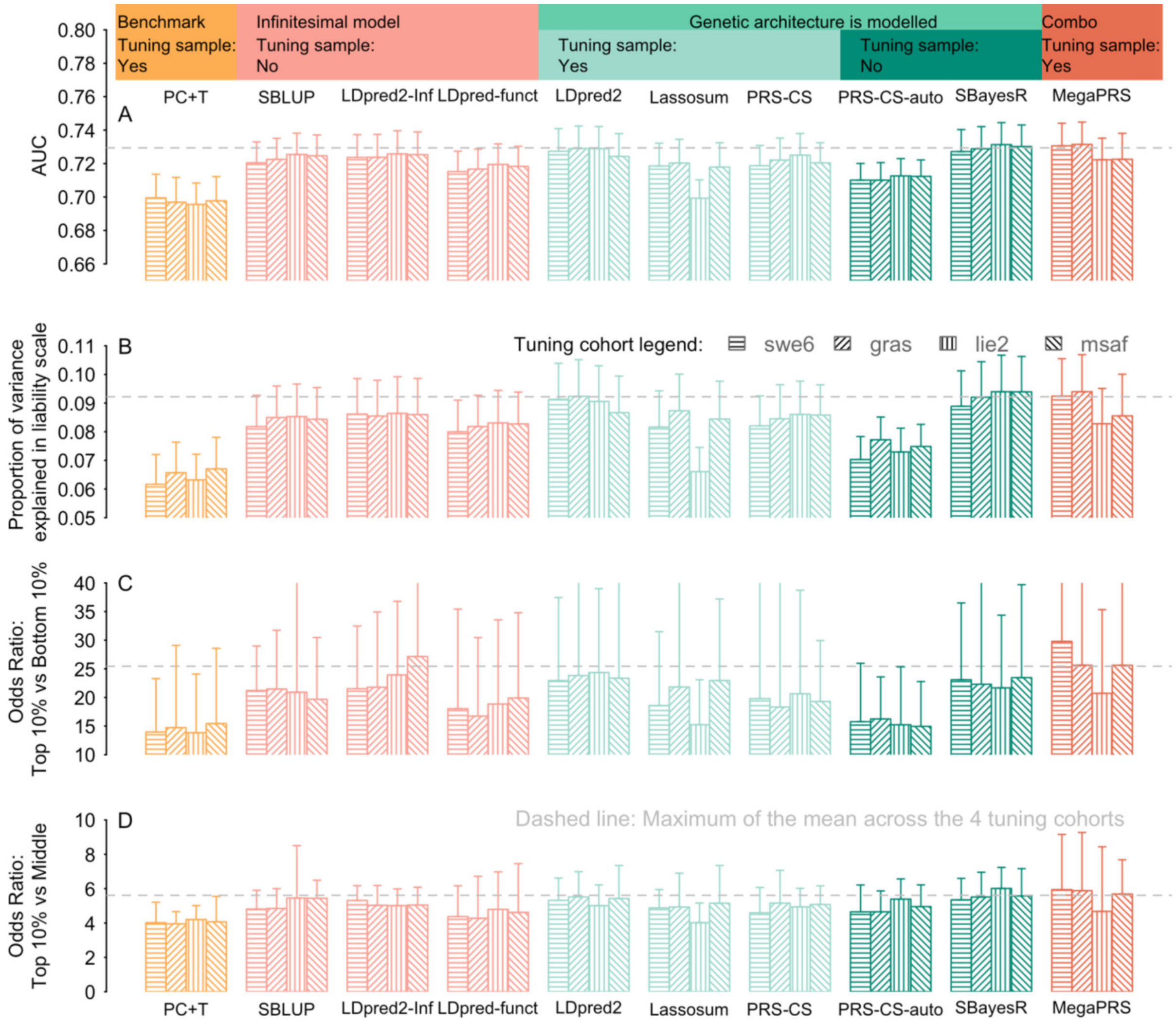

Figure 1. Prediction results for SCZ case/control status using different PGS methods.

The PGS were constructed from SCZ GWAS summary statistics excluding the target cohort and a tuning cohort (shading legend). Each bar reflects the median across 30 target cohorts, the whiskers show the 95% confidence interval for comparing medians. The area under curve (AUC) statistic (A) can be interpreted as the probability that a case ranks higher than a control. Panel (B) is the proportion of variance explained by PGS on the scale of liability, assuming a population lifetime risk of 1%. The third panel (C) is the odds ratio when considering the odds of being a case comparing the top 10% vs bottom 10% of PGS. The bottom panel (D) is the odds of being a case in the top 10% of PGS vs 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 PGS relative to the 5th decile and 6th decile. PC+T (also known as P+T) is the benchmark method which is shown in orange. Pink shows the methods that use an infinitesimal model assumption. The green shows the methods that model the genetic architecture, with light green for the methods using a tuning cohort to determine the genetic architecture of a trait; dark green shows the methods learning the genetic architecture from discovery sample, without using a tuning cohort. Dark orange is for

MegaPRS using the BLD-LDAK model that assume the distribution of SNP effect depends on its allele frequency, LD and function annotation. MegaPRS assign four priors to each of SNP: LASSO, Bridge, BOLT-LMM, BayesR. Each prior has different hyperparameters that identified using the tuning cohort. The dashed grey lines are the maximum of the average across the four 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; gras: 1086 cases, 1232 controls.

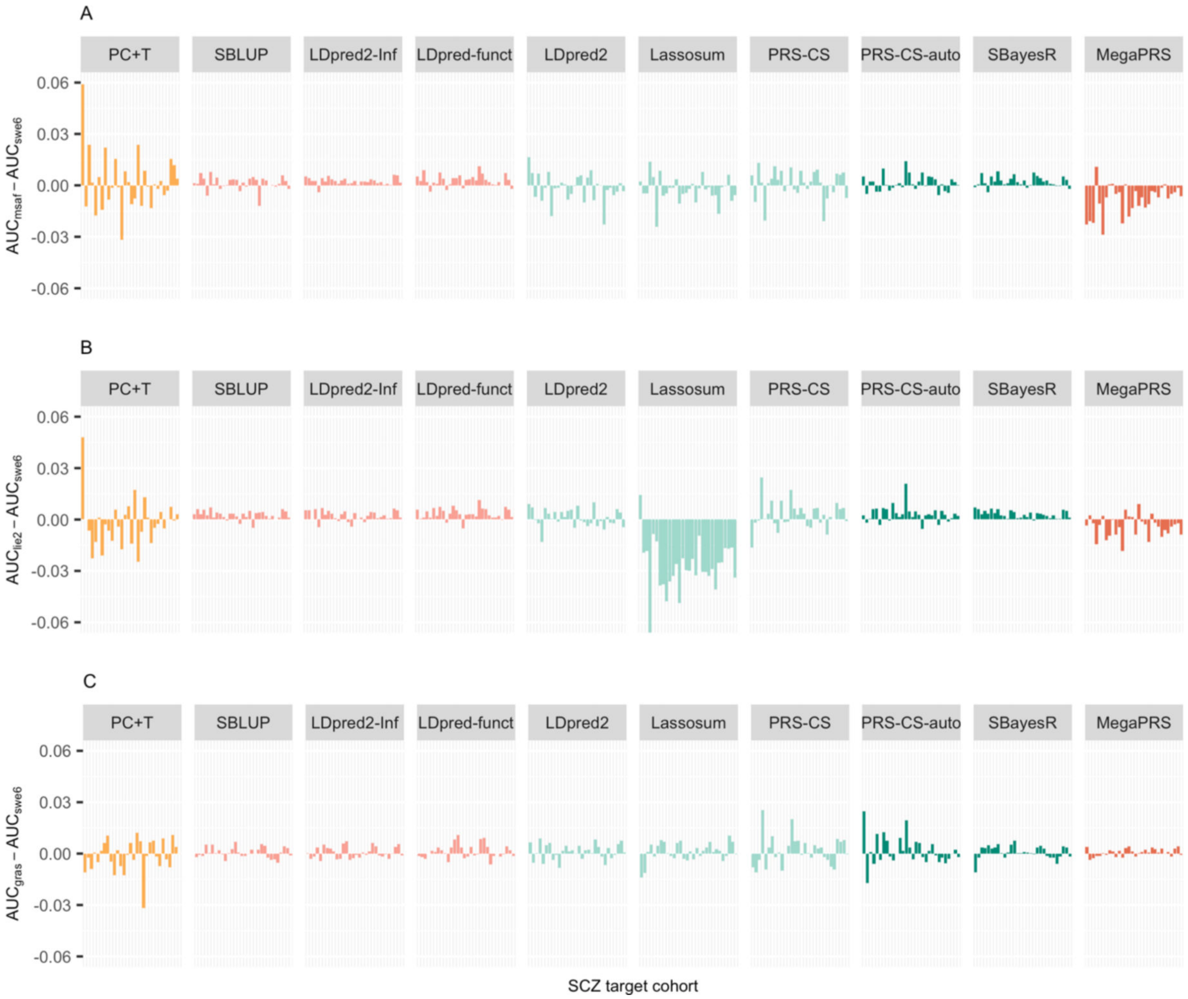

Figure 2. Sensitivity analyses using different tuning cohorts comparing different PGS methods.

Differences in the AUC of SCZ 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. lie2 (137 cases, 269 controls), msaf (327 cases, 139 controls), or gras (1086 cases, 1232 controls)), compared to 'swe6' (1094 cases, 1219 controls). The MAF QC 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 GWAS also changes.

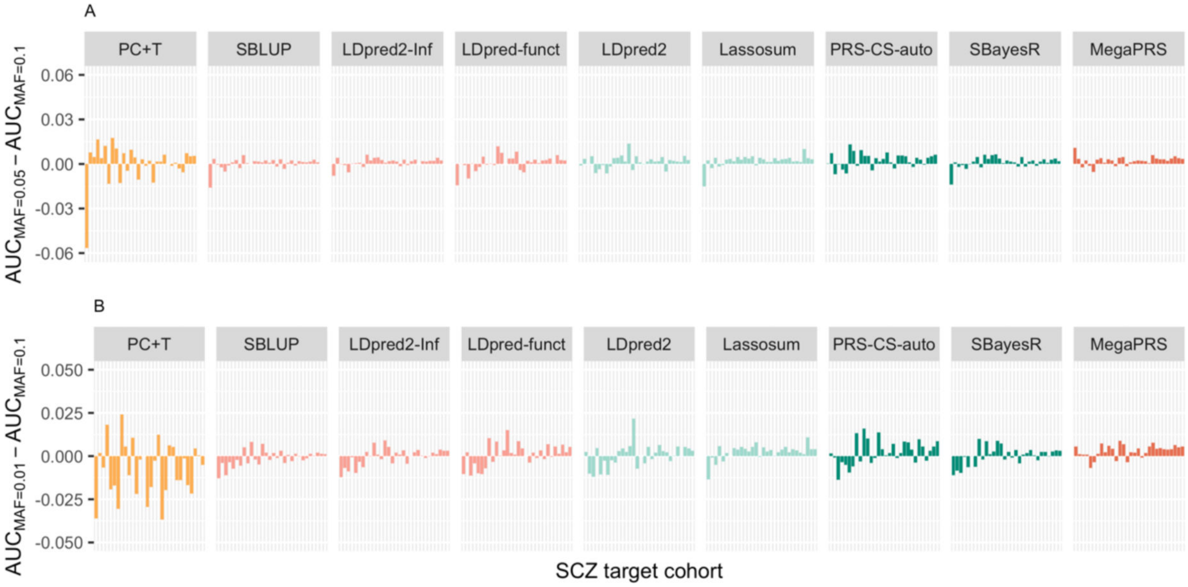

Figure 3. Sensitivity analyses using different MAF quality control thresholds.

Differences in AUC of SCZ 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 A) MAF<0.05 and MAF <0.1 B) MAF<0.01 and MAF <0.1 as a QC threshold. The tuning cohort is 'swe6'.

Table 1.

Summary of methods used to generate PGS

Method	Distribution of SNP effects (β)	Tuning sample	Pre-defined parameters	Parameters estimated in tuning sample
PC+T	None	Yes	-	P value threshold
SBLUP	$\beta \sim N(0, \frac{h_g^2}{m})$ h_g^2: SNP-based heritability, *m*: number of SNPs; $\lambda = m\left(1 - h_g^2\right)/h_g^2$	No	λ LD radius in kb	-
Ldpred2-Inf	Same as SBLUP	No	h_g^2, LD radius in cM or kb	-
LDpred-funct	$\beta_j \sim N\left(0,\ c\sigma_j^2\right)$ $\sum_{j=1}^{M} 1_{\sigma_j^2 > 0} c\sigma_j^2 = h_g^2$, 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 the SNPs is the (1-π) partition are all set to zero.	Yes	h_g^2, π software default values, LD radius in cM or kb	π, sparsity
Lassosum	$f(\boldsymbol{\beta}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} + (1 - s)\ \boldsymbol{\beta}^{\mathrm{T}}\ \mathbf{X}_{\mathrm{r}}^{\mathrm{T}}\mathbf{X}_{\mathbf{r}}\ \boldsymbol{\beta}\ - 2\ \boldsymbol{\beta}^{\mathrm{T}}\ \mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{s}\ \boldsymbol{\beta}^{\mathrm{T}}\ \ \boldsymbol{\beta}\ + 2\lambda\lVert\beta\rVert_1^1$ $\mathbf{X}_r$: genotype of LD reference	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. n is sample size	Yes	a=1, b=0.5 Sample size LD Blocks	ϕ
PRS-CS-auto	Same as PRS-CS, but estimates ϕ from the discovery GWAS.	No	a=1, b=0.5 Sample size LD Blocks	-

Method	Distribution of SNP effects (β)	Tuning sample	Pre-defined parameters	Parameters estimated in tuning sample
SBayesR	$\beta_j \mid \pi, \sigma_\beta^2 \sim \begin{cases} 0, & \text{with probability of } \pi_1 \\ N\left(0, \gamma_2 \sigma_\beta^2\right), & \text{with probability of } \pi_2 \\ \vdots & \\ N\left(0, \gamma_c \sigma_\beta^2\right), & \text{with probability of } 1 - \sum_{c=1}^{C-1} \pi_c \end{cases}$ $\sigma_\beta^2 \sim Inv - \chi^2(d.f. = 4)$ $\boldsymbol{\pi}_i \sim Dir(\mathbf{1})$, 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: $\beta_j \sim N\left(0, \upsilon\sigma_j^2\right)$ BOLT-LMM: $\beta_j \sim \begin{cases} N\left(0, (1 - f_2)/\pi\sigma_j^2\right), & \text{with probability of } \pi \\ N\left(0, (f_2)/(1 - \pi)\sigma_j^2\right), & \text{with probability of } 1 - \pi \end{cases}$ f_2 is the proportion of the total mixture variance in the second normal distribution. BayesR: similar to SBayesR with C=4, and $\boldsymbol{\pi}_i$ 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; $\lVert \boldsymbol{\beta} \rVert_1^1 = \sum_i \lvert\beta_i\rvert$. When h_g^2 (SNP-based heritability) is a pre-defined parameter it is estimated from the discovery GWAS, where “discovery GWAS” is the genome-wide set of association statistics (SNP ID, reference allele, frequency of reference allele, association effect size for reference allele, standard error of effect size, association p-value, sample size). We use bold for matrix notation and italics for scalar notation. cM: centimorgan; kb: kilobase pair