Evaluation of Polygenic Prediction Methodology within a Reference-Standardized Framework

Oliver Pain[1,2], Kylie P. Glanville[1], Saskia Hagenaars[1], Saskia Selzam[1], Anna E. Fürtjes[1], Helena Gaspar[1], Jonathan R. I. Coleman[1], Kaili Rimfeld[1], Gerome Breen[1,2], Robert Plomin[1], Lasse Folkersen[3], Cathryn M. Lewis[1,2,4].

[1]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, SE5 8AF, United Kingdom
[2]NIHR Maudsley Biomedical Research Centre, South London and Maudsley NHS Trust, London, SE5 8AF, UK.
[3]Institute of Biological Psychiatry, Sankt Hans Hospital, Copenhagen, 4000 Roskilde, Denmark
[4]Department of Medical and Molecular Genetics, Faculty of Life Sciences and Medicine, King's College London, London, SE1 9RT, UK.

Corresponding author:
Dr. Oliver Pain
Social, Genetic and Developmental Psychiatry Centre
Institute of Psychiatry, Psychology and Neuroscience - PO80
De Crespigny Park, Denmark Hill, London,
United Kingdom, SE5 8AF
Phone: 02078485299
Email: oliver.pain@kcl.ac.uk

Running title: Polygenic prediction comparison
Keywords: polygenic scores, reference-standardized, prediction, elastic net, GWAS

Abstract

Background: The predictive utility of polygenic scores is increasing, and many polygenic scoring methods are available, but it is unclear which method performs best. This study evaluates the predictive utility of polygenic scoring methods within a reference-standardized framework, which uses a common set of variants and reference-based estimates of linkage disequilibrium and allele frequencies to construct scores.

Methods: Six polygenic score methods were tested: p-value thresholding and clumping (pT+clump), SBLUP, lassosum, LDPred, PRScs and SBayesR, evaluating their performance to predict outcomes in UK Biobank and the Twins Early Development Study (TEDS). Strategies to identify optimal p-value threshold and shrinkage parameters were compared, including 10-fold cross validation, pseudovalidation (with no validation sample), and multi-polygenic score elastic net models.

Results: lassosum, PRScs and LDPred performed strongly using 10-fold cross-validation to identify the most predictive p-value threshold or shrinkage parameter, giving a relative improvement of 14-17% over pT+clump in the correlation between observed and predicted outcome values. Using pseudovalidation, the best method was PRScs, with a relative improvement of >11% over other pseudovalidation methods (lassosum, SBLUP, SBayesR, LDPred), and only 1% less than the best polygenic score identified by 10-fold cross validation. Elastic net models containing polygenic scores based on a range of parameters consistently improved prediction over any single polygenic score.

Conclusion: Within a reference-standardized framework, the best polygenic prediction was achieved using lassosum and PRScs, modeling multiple polygenic scores derived using multiple parameters. This study will help researchers performing polygenic score studies to select the most powerful and predictive analysis methods.

Introduction

In personalized medicine, medical care is tailored for the individual to provide improved disease prevention, prognosis, and treatment. Genetics is a potentially powerful tool for providing personalized medicine as genetic variation accounts for a large proportion of individual differences in health and disease [1]. Furthermore, an individual's genetic sequence is stable across the lifespan, enabling predictions long before the onset of most diseases. Although genetic information is used to predict rare Mendelian genetic disorders, our ability to predict common disorders using genetic information is currently insufficient for clinical implementation. This is due to the increased etiological complexity of common disorders, with complex interplay between genetic and environmental factors, and the highly polygenic genetic architecture with contributions from many genetic variants with small effect sizes [2]. However, genome-wide association studies (GWAS), used to detect common genetic associations, are rapidly increasing in sample size, and are identifying large numbers of novel and robust genetic associations for health-related outcomes [3]. This growing source of information is also improving our ability to predict an individual's disease risk or measured trait based on their genetic variation [4,5].

An individual's genetic risk for an outcome can be summarized in a polygenic score, calculated from the number of trait-associated alleles carried. The contributing variants are typically weighted by the magnitude of effect they confer on the outcome of interest, estimated in a reference GWAS. There are several challenges in performing a well-powered polygenic score analysis. Firstly, GWAS effect-sizes are inflated through Winner's curse, and unbiased estimates can only be obtained through an independent training sample, with these effect-size estimates then used to calculate polygenic scores in a further independent sample [6]. Secondly, to maximize polygenic prediction accuracy, the GWAS summary statistics must be adjusted to account for the linkage disequilibrium (LD) between genetic variants, to avoid double counting the non-independent effect of variants in high LD, and account for varying degrees of polygenicity across outcomes, i.e. the number of genetic variants affecting the outcome [6]. LD can be accounted for using LD-based clumping of GWAS summary statistics, removing variants in LD with lead variants within each locus, and polygenicity is accounted for by applying multiple GWAS *p*-value thresholds (pT) to select the effect alleles included in the polygenic score [4,5]. This pT+clump approach is conceptually simple and computationally scalable [7]. However, using a hard LD threshold in clumping to retain or remove variants from the polygenic score calculation can potentially reduce the variance explained by the polygenic score. Alternative summary statistic-based polygenic score methods retain all genetic variants by modelling both the LD between variants and the polygenicity of the outcome [8–12]. These methods use estimates of LD to jointly estimate the effect of nearby genetic variation maximizing the signal captured, and generally apply a shrinkage parameter to the genetic effects to reduce overfitting and allow for varying degrees of polygenicity across outcomes.

Effect sizes estimated in a GWAS are typically larger than they would be in an independent sample due to overfitting or winner's curse. Effect size estimates can be reduced using shrinkage methods to improve the generalizability of the model. Shrinkage methods for polygenic scoring can be separated into frequentist penalty-based methods (e.g. lasso regression-based lassosum [10], summary-based best linear unbiased prediction (SBLUP) [9]) and Bayesian methods that shrink estimates to fit a prior distribution of effect sizes, such as LDPred [8], PRScs [11] and SBayesR [12]. Each of these methods has been shown to improve the predictive utility of polygenic scores over those derived using the pT+clump approach. In comparisons between methods the findings are mixed: some studies have similar results across methods [13], while papers developing a new method often report that the developed method out-performs chosen other methods. To our knowledge no independent study has yet compared all approaches.

Four methods (pT+clump, LDPred, lassosum and PRScs) generate multiple polygenic scores from user-defined tuning parameters. To determine which tuning parameter provides optimal prediction, the polygenic scores must first be tested in an independent 'tuning' sample. The pT+clump approach applies p-value thresholds to select variants included in the polygenic score, whereas LDPred, lassosum and PRScs apply shrinkage parameters to adjust the GWAS effect sizes. In addition, lassosum and PRScs also provide a pseudovalidation approach, whereby a single optimal shrinkage parameter is estimated based on the GWAS summary statistics alone, and therefore do not require a tuning sample. Two further methods, SBLUP and SBayesR, can be considered pseudovalidation approaches as they also do not require a tuning sample to identify optimal parameters. Rather than selecting a single tuning parameter, some studies have suggested that combining polygenic scores across p-value thresholds whilst taking into account their correlation using either PCA or model stacking can improve prediction [14,15].

Polygenic scores are a useful research tool, as well as a promising potential tool for personalized healthcare through prediction of disease risk, prognosis, and treatment response [16]. However, polygenic scores calculated in a clinical setting should be valid for a single target sample and thus need to be constructed using a reference-standardized framework. Here, the polygenic score is independent of any properties specific to the target sample, including the genetic variation available, and the LD and minor allele frequency (MAF) estimates. In a reference-standardized approach, the genetic variants considered can be standardized by using only single nucleotide polymorphisms (SNPs) that are commonly available after imputation, such as variation within the HapMap3 reference [17]. The LD and MAF estimates can be standardized by using an ancestry matched individual-level genetic dataset such as 1000 Genomes [18]. Determining these properties (SNPs, LD, MAF) in reference data provides a practical approach for estimating polygenic scores for an individual, making them comparable to polygenic scores for other individuals of the same ancestry [19]. Use of a reference-standardized framework also offers advantages by improving the comparability of polygenic scores across cohorts. Several polygenic scoring methods now recommend the use of HapMap3 SNPs and precomputed external LD estimate references [11,12], in line with a reference-standardized approach.

In this study, we perform an extensive comparison of polygenic scoring methods within a reference-standardized framework. We evaluate the predictive utility of models for outcomes in UK Biobank and TEDS, combining information across tuning parameters. We evaluate six polygenic scoring methods and apply different modelling strategies to select optimal tuning parameters to establish the combinations that perform consistently well. The reference-standardized framework increases the generalizability of results and provides a resource for future studies investigating polygenic prediction in a research study or clinical setting.

Methods

To evaluate the different polygenic scoring approaches, we used two target samples: UK Biobank (UKB) [20], and the Twins Early Development Study (TEDS) [21]. All code used to prepare data and carryout analyses is available on the GenoPred website (see Data and Code Availability).

UKB

UKB is a prospective cohort study that recruited >500,000 individuals aged between 40-69 years across the United Kingdom. The protocol and consent were approved by the UKB's Research Ethics Committee (Ref: 11/NW/0382).

Genetic data

UKB released imputed dosage data for 488,377 individuals and ~96 million variants, generated using IMPUTE4 software [20] with the Haplotype Reference Consortium reference panel [22] and the UK10K Consortium reference panel [23]. This study retained individuals that were of European ancestry based on 4-means clustering on the first 2 principal components provided by the UKB, and removed related individuals (>3rd degree relative) using relatedness kinship (KING) estimates provided by the UKB [20]. The imputed dosages were converted to hard-call format using a hard call threshold of zero.

Phenotype data

Nine UKB phenotypes were analyzed. Six phenotypes were binary: Depression, Type II Diabetes (T2D), Coronary Artery Disease (CAD), Inflammatory Bowel Disorder (IBD), Rheumatoid arthritis (RheuArth), and Multiple Sclerosis (MultiScler). Three phenotypes were continuous: Intelligence, Height, and Body Mass Index (BMI). Further information regarding outcome definitions can be found in the Supplementary Material.

Analysis was performed on a subset of ~50,000 UKB participants for each outcome. For each continuous trait (Intelligence, Height, BMI), a random sample was selected. For disease traits, all cases were included, except for Depression and CAD where a random sample of 25,000 cases was selected. Controls were randomly selected to obtain a total sample size of 50,000. Sample sizes for each phenotype after genotype data quality control are shown in Table 1.

TEDS

The Twins Early Development Study (TEDS) is a population-based longitudinal study of twins born in England and Wales between 1994 and 1996 [24]. Ethical approval for TEDS has been provided by the King's College London ethics committee (reference: 05/Q0706/228). Parental and/or self-consent was obtained before data collection. For this study, one individual from each twin pair was removed to retain only unrelated individuals.

Genetic data

TEDS participants were genotyped using two arrays, HumanOmniExpressExome-8v1.2 and AffymetrixGeneChip 6.0. Stringent quality control was performed separately for each array, prior to imputation via the Sanger Imputation server using the Haplotype Reference Consortium (release 1.1) reference data [22,25]. Imputed genotype dosages were converted to hard-call format using a hard call threshold of 0.9, with variants for each individual set to missing if no genotype had a probability of >0.9. Variants with an INFO score < 0.4, MAF < 0.001, missingness > 0.05 or Hardy-Weinberg equilibrium p-value < 1×10^{-6} were removed.

Phenotypic data

This study used four continuous phenotypes within TEDS: Height, Body Mass Index (BMI), Educational Achievement, and Attention Deficit Hyperactivity Disorder (ADHD) symptom score (Table 1). These phenotypes were selected based on a previous polygenic study, enabling comparison across methods [26]. The phenotypes were derived using the same protocol as previously.

Genotype-based Scoring

The following genotype-based scoring procedure provides reference standardized polygenic scores and can be applied to any datasets of imputed genome-wide array data (Figure 1).

SNP-level QC

HapMap3 variants from the LD-score regression website (see Web Resources) were extracted from target samples (UKB, TEDS), inserting any HapMap3 variants that were not available in the target sample as missing genotypes (as required for reference MAF imputation by the PLINK allelic scoring function) [27]. No other SNP-level QC was performed.

Individual-level QC

Individual-level QC prior to imputation was previously performed for both UKB [20] and TEDS [25] samples. Only individuals of European ancestry were retained for polygenic score analysis. They were identified using 1000 Genomes Phase 3 projected principal components of population structure, retaining only those within three standard deviations from the mean for the top 100 principal components. This process will also remove individuals who are outliers due to technical genotyping or imputation errors.

GWAS summary statistics

GWAS summary statistics were identified for phenotypes the same as or similar as possible to the UKB and TEDS phenotypes (descriptive statistics in Table S1), excluding GWAS with documented sample overlap with the target samples. GWAS summary statistics were formatted using the LD-Score Regression munge_sumstats.py script (see Web Resources) with default settings except the minimum INFO score was set to 0.6. The munged GWAS summary statistic files only contain information on the SNP ID, reference and effect allele, Z-score of association, and the sample size per variant. Note, per variant sample size was only available for seven of the 12 GWAS used (Table S1).

Reference genotype datasets

Target sample genotype-based scoring was performed using two different reference genotype datasets, the European subset of 1000 Genomes Phase 3 (N=503) and a random subset of 10,000 European-ancestry UKB participants. The UKB reference set was independent of the target sample used for evaluating polygenic scoring methods. These references were used to determine whether the sample size of the reference genotype dataset affects the prediction accuracy of polygenic scores. Only 1,042,377 HapMap3 variants were available in the UKB dataset and used in genotype-based scoring.

Polygenic Scores (PRS)

Polygenic scoring was carried out using six approaches outlined in Table 2. To ensure comparability across methods, the same set of HapMap3 variants were considered, and the same reference genotype datasets were used to estimate LD and MAF (except for PRScs and SBayesR).

PRScs-provides an LD reference for HapMap3 variants based on the European subset of the 1000 Genomes, and results should be comparable to other methods when using the 1000 Genomes reference. PRScs was not applied using the larger UKB reference dataset as PRScs has been previously reported to show minimal improvement when using larger LD reference datasets [11].

SBayesR analysis requires shrunk and sparse LD matrices as input. LD matrices were calculated using Genome-wide Complex Trait Bayesian analysis (GCTB) [28] in batches of 5,000 variants, which were then merged for each chromosome, shrunk, and then made sparse. SBayesR analysis was also performed using LD matrices released by the developers of GCTB based on 50,000 European UKB individuals (see Web Resources).

Two additional modifications of the standard pT+clump approach were tested, termed 'pT+clump (non-nested)' and 'pT+clump (dense)'. The pT+clump (non-nested) approach is the same the standard pT+clump approach except non-overlapping p-value thresholds were used to select variants included in the polygenic score, thereby making the polygenic scores for each threshold independent. The pT+clump (dense) approach is the same as the standard pT+clump approach except that it uses 10,000 p-value thresholds (minimum=5×10^{-8}, maximum=0.5, interval=5×10-5), implemented using default settings in PRSice [7].

After adjustment of GWAS summary statistics as necessary for each polygenic scoring method, polygenic scores were calculated using PLINK with reference MAF imputation of missing data. All scores were standardized based on the mean and standard deviation of polygenic scores in the reference sample.

Modelling approaches

For methods that provide polygenic scores based on a range of p-value thresholds (pT+clump) or shrinkage parameters (lassosum, PRScs, LDPred), the best parameter was identified using either 10-fold cross validation (10FCVal) and, if available, pseudovalidation (PseudoVal). Pseudovalidation was performed using the pseudovalidate function in lassosum, the fully-Bayesian approach in PRScs, and the infinitesimal model in LDPred. In addition to selecting the single 'best' parameter for polygenic scoring, elastic net models were derived containing polygenic scores based on a range of parameters, with elastic net shrinkage parameters derived using 10-fold cross-validation (Multi-PRS). SBLUP and SBayesR methods both use only a single shrinkage parameter, and therefore considered pseudovalidation approaches.

The optimal parameters (pT, GWAS-effect size shrinkage, elastic net parameters) were determined based on the largest mean correlation between observed and predicted values obtained through 10-fold cross validation, and the resulting model was then applied to an independent test set. Ten-fold cross-validation is liable to overfitting when using penalized regression as hyperparameters are tuned using the 10-fold cross validation procedure. The independent test-set validation avoids any overfitting as the independent test sample is not used for hyperparameter tuning. Ten-fold cross validation was performed using 80% of the sample and the remaining 20% was used as the independent test sample. Ten-fold cross validation and test-set validation was carried out using the 'caret' R package, setting the same random seeds prior to subsetting individuals to ensure the same individuals were included for all polygenic scoring methods.

Evaluating prediction accuracy

Prediction accuracy was evaluated as the Pearson correlation between the observed and predicted outcome values. Correlation was used as the main test statistic as it is applicable for both binary and continuous outcomes and standard errors are easily computed as

$$SE_r = \frac{1 - r^2}{\sqrt{n - 2}} \qquad (1)$$

Where SE_r is the standard error of the Pearson correlation, r is the Pearson correlation, and n is the sample size. Correlations can be easily converted to other test statistics such as R^2 (observed or liability) and area under the curve (AUC) (equations 8 and 11 in [29]), with relative performance of each method remaining unchanged.

Logistic regression was used for predicting binary outcomes, and linear regression was used for predicting continuous outcomes. If the model contained only one predictor, a generalized linear model was used. If the model contained more than one predictor (i.e. the polygenic scores for each p-value threshold or shrinkage parameter), an elastic net model was applied to avoid overfitting due to the inclusion of multiple correlated predictors [30].

The correlation between observed and predicted values of each model were compared using William's test (also known as the Hotelling-Williams test) [31] as implemented by the 'psych' R package's 'paired.r' function, with the correlation between model predictions of each method specified to account for their non-independence. A two-sided test was used when calculating p-values.

The correlation between predicted and observed values were combined across phenotypes for each polygenic score method. Correlations and their variances (SE^2) were aggregated using the 'BHHR' method [32] as implemented in the 'MAd' R package's 'agg' function, using a phenotypic correlation matrix to account for the non-independence of analyses within each target sample.

The percentage difference between methods was calculated as

$$\%\ difference\ = ((r_1 - r_2)/r_2) * 100 \qquad (2)$$

Where r_1 and r_2 indicate the Pearson correlation between predicted and observed values for models 1 and 2, respectively.

Results

The six polygenic risk score methods were applied to the target datasets of UKB (9 phenotypes) and TEDS (4 phenotypes), using two reference data sets of 1000 Genomes (1KG, 503 individuals) and UKB (10,000 individuals). Models were derived using 10-fold cross-validation, pseudovalidation, and analysis of multiple threshold PRS, as appropriate for each polygenic risk score method (Table 2).

First, we confirmed that the design of the study was appropriate to detect differences between the methods using the GWAS summary statistics and test data sets chosen. GWAS summary statistics had sample sizes of a mean of 39,807 cases and 110,649 controls, and 229,313 individuals for continuous traits, with heritability on the liability scale (estimated from the GWAS) ranging between 0.021 (Multiple Sclerosis) and 0.542 for Crohn's disease (Table S1). For pT+clump, with UKB reference and target panel, the correlations between observed values (case-control status or measured trait) and the predicted values from the polygenic risk scoring models ranged from 0.069 (SE=0.010) for Multiple Sclerosis to 0.297 (SE=0.010) for Height (Table S7). For each disorder or trait, reference panel and polygenic scoring method, the correlation was significantly different from zero (Tables S6-S9). These results confirm that the study design - comprising the GWAS, reference panel, target studies and traits - had sufficient information to capture polygenic prediction, and that the traits are diverse in polygenic architecture.

Results were highly concordant across the different target and reference samples used though the estimates were more precise when using the UKB target sample due to the increased sample size compared to TEDS (Figure S1-S2).

Effect of reference panel and validation method

All polygenic scoring methods were applied to two reference panels of European ancestry: 503 individuals from the 1,000 Genomes sample, and 10,000 individuals from UKB. Results were highly similar for both panels (Figure S1-S2). For example, with the larger reference panel the correlation increased by a mean of 0.0054 in UKB, and 0.0040 in TEDS, across traits and polygenic scoring methods (test-set validation, Table S2-S5; excluding PRScs which used only the 1,000 Genomes reference panel). Detailed results are reported here only for the 1,000 Genomes (1KG) reference panel, with full results for UKB reference panel in Supplementary Materials.

Both 10-fold cross validation and test-set validation methods were used in modelling, across all polygenic risk scoring methods. The 10-fold cross validation results were highly congruent with test-set validation results (Table 3). Results reported are based on test-set validation since this method is clearly robust to overfitting when using elastic net models (see Supplementary Materials for 10-fold cross-validation results).

Overview of polygenic scoring methods by modelling strategy

The performance for each polygenic scoring method across phenotypes was assessed using the correlation between observed and fitted values (Figure 2A), and then comparing each method with a baseline method of pT+clump with 10-fold cross validation using the difference in correlation (Figure 2B). All methods performed at least as well as pT+clump, except for SBayesR, which had convergence problems for several of the phenotypes (see Supplementary Material for full information). These overview results show that for the methods where all three modelling strategies could be performed (lassosum, PRScs, and LDPred), pseudovalidation performed less well, and that the prediction when modelling multiple PRS was slightly higher than the 10-fold cross. Full results for all traits in UKB and TEDS indicate consistency across methods, with no trait performing unexpectedly well or poorly on any single method (Tables S6-S9; Figures S3-S6).

Comparison of polygenic scoring methods

A pairwise comparison of polygenic scoring methods was performed for each method (pT+clump, lassosum, PRScs, SBLUP, SBayesR, LDPred) and each model (10-fold cross validation, multiPRS and pseudovalidation). Figure 3 shows the difference in correlation (R) within and between methods for UKB outcomes with 1KG reference panel, with p-values for significant differences calculated using the William's test results aggregated across outcomes. Full results for TEDS and UKB, and for both reference panels are given in Tables S10-S13, and by trait in Tables S14-S17.

When using 10-fold cross validation to identify the optimal parameter, lassosum, PRScs and LDPred provided the most predictive polygenic scores in the test sample on average, with a 14-17% relative improvement ($p<3\times10^{-9}$) over the 10-fold cross-validated pT+clump approach.

Of the methods providing a pseudovalidation approach (lassosum, PRScs, LDPred, SBLUP and SBayesR), PRScs performed the best on average, providing at least an 11% relative improvement ($p<2\times10^{-11}$) over other pseudovalidation approaches. Furthermore, the PRScs pseudovalidation approach was on average only 1% (*p*-value = 0.19) worse than the best polygenic score identified by 10-fold cross validation for any method. The performance of lassosum pseudovalidation, the LDPred infinitesimal model, SBLUP and SBayesR was variable across phenotypes, whereas the PRScs pseudovalidated polygenic score achieved near optimal predication compared to any method. The pseudovalidated PRScs polygenic score was only significantly improved upon by the elastic-net model containing multiple lassosum (6% relative improvement, $p=2.69\times10^{-8}$) and PRScs polygenic scores (4% relative improvement, $p=2.23\times10^{-8}$).

Modelling multiple polygenic scores based on multiple parameters using an elastic net consistently outperformed models containing the single best polygenic score as identified using 10-fold cross validation. The improvement was largest when using pT+clump polygenic scores (13% relative improvement, $p=9.78\times10^{-19}$), but was also statistically significant for lassosum (6% relative improvement, 6.59×10^{-12}), PRScs (3% relative improvement, $p=5.56\times10^{-5}$), and LDPred (1% relative improvement, $p<6.72\times10^{-3}$) methods. Elastic net models using non-nested or dense *p*-value thresholds showed no improvement over the standard *p*-value thresholding approach (Tables S18-S19).

The performance of SBayesR was higher when using the larger UKB reference sample (Figures S1-S2, S11-S12), though on average it still performed worse than all other approaches (Figure S7). SBayesR results based on the UKB reference were similar to those using the GCTB LD reference (Figures S11-S12). The relative performance of SBayesR varied substantially (Figures 2B, S2, S8-S9). The SBayesR SNP-based heritability for the Height GWAS was >1 (indicating poor convergence), even when restricting variants to P<0.4 as suggested by the methods developers (Table S20). The Height performance was included in the average results to convey this limitation of SBayesR. The SBayesR heritability results for each GWAS when using different approaches for preparing the summary statistics are shown in Table S20.

Discussion

This study evaluated a range of polygenic scoring methods across phenotypes representing a range of genetic architectures and using reference and target sample genotypic data of different sample sizes. This study shows that, when a tuning sample is available to identify optimal parameters, more recently developed methods that do not perform LD-based clumping provide better prediction, with lassosum, PRScs and LDPred providing a relative improvement of 14-17% compared to the pT+clump approach. When a tuning sample is not available, the optimal method for prediction was PRScs, with its pseudovalidated polygenic score providing an >11% relative improvement over other pseudovalidation approaches. Furthermore, the PRScs pseudovalidation performance was not significantly worse than the best polygenic scores identified by 10-fold cross validation for any other method. This study also shows that an elastic net model containing multiple polygenic scores based on a range of p-value thresholds or shrinkage parameters provides better prediction than the single best polygenic score as identified by 10-fold cross validation. Modelling multiple parameters increased prediction by 13% when using the pT+clump approach and 1-6% for polygenic scoring methods that model LD.

These methods were evaluated within a reference-standardized framework and the results are likely to be generalizable to a range of settings, including a clinical setting. The improved transferability of prediction accuracy when using a reference-standardized approach enables prediction with a known accuracy for a single individual. This is an essential feature of any predictor as then its prediction can be appropriately considered in relation to other information about the individual. It is important to consider whether the reference-standardized approach impacts the predictive utility of the polygenic scores compared to those derived using target sample specific properties. The use of only HapMap3 variants is common for polygenic scoring methods as denser sets of variants increase the computational burden of the analysis and provide only incremental improvements in prediction [12]. However, denser sets of variants are ultimately likely to be of importance for optimizing the predictive utility of polygenic scores. The use of reference LD estimates instead of target sample-specific LD estimates is less likely to impact the predictive utility of polygenic scores. LD estimates are used to recapitulate LD structure in the GWAS discovery sample, and there should therefore be no advantage to using target sample specific LD estimates instead of reference sample LD estimates, unless the target sample better captures the LD structure in the GWAS discovery sample.

One major limitation of our study is that it was performed only in studies of European ancestry since GWAS of other ancestries have insufficient power for polygenic prediction. Polygenic scoring method comparisons in other ancestries or across ancestries will require substantial progress in diversifying genetic studies to non-European ancestry. In particular, it will be important to assess the impact of greater genetic diversity and weaker linkage disequilibrium in African ancestry populations. These studies are essential if polygenic risk scores are to be implemented in clinical care, to ensure equity of healthcare.

The clinical implementation of polygenic scores is at an early stage, and we identify five areas that still require further research. First, the influence of accounting for principal components of ancestry should be explored with regard to a range of polygenic scoring approaches. Previous research has shown that when prediction is the aim, and inference is not of interest, adjusting polygenic scores for their relationship with principal components of ancestry reduces their predictive utility [33]. Instead, including principal components as independent predictors in the prediction model can improve prediction over polygenic scores alone [33]. The extent to which integration of principal components of ancestry or local ancestry can improve prediction with regard to a range of polygenic scoring methods is yet to be fully investigated systematically. Second, this study demonstrates that the reference-standardized approach provides reliable polygenic score estimates. However, the extent to which missing genetic variation within target sample data affects the prediction accuracy needs to be investigated. Furthermore, the extent to which prediction accuracy varies across individuals from different European ancestral populations needs to be assessed. Third, this study used the HapMap3 SNP list when deriving polygenic scores, building on previous research suggesting that these variants are reliably imputed and provide good coverage of the genome [17]. However, other sets of variants should be explored as denser coverage of the genome may improve prediction. Fourth, this study investigates polygenic scores based on a single discovery GWAS or phenotype. Previous research has shown that methods which combine evidence across multiple GWAS can improve prediction due to genetic correlation between traits [34–37]. Further research comparing the predictive utility of multi-trait polygenic prediction within a reference-standardized framework is required. Finally, integration of functional genomic annotations has been shown to improve prediction over functionally-agnostic polygenic scoring methods [38]. Comparison of functionally-informed methods within a reference-standardized framework is also required.

In conclusion, this study performed a comprehensive comparison of GWAS summary statistic-based polygenic scoring methods within a reference-standardized framework using European ancestry studies. The results provide a useful resource for future research and endeavors to implement polygenic scores for individual-level prediction. All the code, rationale and results of this study are available on the GenoPred website (see Web Resources). This website will continue to document the evaluation of novel genotype-based prediction methods, providing a valuable community resource for education, research, and collaboration. Novel polygenic score methods can be rapidly tested against these standard methods to benchmark performance. This framework should be a valuable tool in the roadmap of moving polygenic risk scores from research studies to clinical implementation. Further investigation of methods providing genotype-based prediction within a reference-standardized framework is needed.

Declaration of Interests

Cathryn Lewis sits on the Myriad Neuroscience Scientific Advisory Board. The other authors declare no competing interests.

Acknowledgements

We thank Luke Lloyd-Jones and Jian Zeng for their advice on SBayesR using GCTB, and Tian Ge on using PRScs. We thank Paul O'Reilly and Sam Choi for useful discussion.

This paper represents independent research funded by the UK Medical Research Council (MR/N015746/1 and MR/S0151132), and the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The authors acknowledge use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk), which is delivered in partnership with the NIHR Maudsley BRC, and part-funded by capital equipment grants from the Maudsley Charity (award 980) and Guy's & St. Thomas' Charity (TR130505). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. We thank the research participants and employees of 23andMe for making the work regarding Depression possible.

UKB: This research was conducted under UK Biobank application 18177.

TEDS: We gratefully acknowledge the ongoing contribution of the participants in TEDS and their families. TEDS is supported by UK Medical Research Council Program Grant MR/M021475/1 (and previously Grant G0901245) (to Robert Plomin (R.P).), with additional support from National Institutes of Health Grant AG046938. The research leading to these results has also received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC Grant Agreement 295366. R.P. is supported by Medical Research Council Research Professor Award G19/2. KR is supported by a Sir Henry Wellcome Postdoctoral Fellowship.

Web Resources

- LDSC HapMap 3 SNP-list: https://data.broadinstitute.org/alkesgroup/LDSCORE/w_hm3.snplist.bz2
- LDSC Munge Sumstats: https://github.com/bulik/ldsc/blob/master/munge_sumstats.py
- GCTB LD matrices: https://zenodo.org/record/3350914
- Impute.me: https://impute.me/
- GenoPred: https://opain.github.io/GenoPred

Data and Code Availability

The code used during this study are available at GitHub: https://opain.github.io/GenoPred. An application is required to access individual-level data for TEDS and UKB.

References

1. Polderman, T.J.C., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher, P.M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat. Genet. *47*, 702–709.
2. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., and Chakravarti, A. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
3. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., and Sollis, E. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.
4. Consortium, I.S. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748.
5. Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet. *9*,.
6. Choi, S.W., Mak, T.S.H., and O'reilly, P. (2018). A guide to performing Polygenic Risk Score analyses. BioRxiv 416545.
7. Choi, S.W., and O'Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience *8*, giz082.
8. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., and Do, R. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592.
9. Robinson, M.R., Kleinman, A., Graff, M., Vinkhuyzen, A.A.E., Couper, D., Miller, M.B., Peyrot, W.J., Abdellaoui, A., Zietsch, B.P., and Nolte, I.M. (2017). Genetic evidence of assortative mating in humans. Nat. Hum. Behav. *1*, 16.
10. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., and Sham, P.C. (2017). Polygenic scores via penalized regression on summary statistics. Genet. Epidemiol. *41*, 469–480.
11. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1–10.
12. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., and Esko, T. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. *10*, 1–11.
13. Allegrini, A.G., Selzam, S., Rimfeld, K., von Stumm, S., Pingault, J.-B., and Plomin, R. (2019). Genomic prediction of cognitive traits in childhood and adolescence. Mol. Psychiatry *24*, 819–827.
14. Coombes, B.J., and Biernacka, J.M. (2019). A principal component approach to improve association testing with polygenic risk scores. BioRxiv 847020.
15. Privé, F., Vilhjálmsson, B.J., Aschard, H., and Blum, M.G.B. (2019). Making the most of Clumping and Thresholding for polygenic scores. Am. J. Hum. Genet. *105*, 1213–1221.
16. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. Genome Med. *12*, 1–11.
17. International HapMap 3 Consortium, investigators, P., leaders, P. coordination, group, M. writing, QC, G., and discovery, E. 3 sequencing, SNP, analysis, C. number variation typing, and analysis, P., analysis, L. frequency variation, et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
18. Consortium, 1000 Genomes Project (2015). A global reference for human genetic

variation. Nature *526*, 68–74.
19. Folkersen, L., Pain, O., Ingason, A., Werge, T., Lewis, C.M., and Austin, J. (2019). Impute. me: an open source, non-profit tool for using data from DTC genetic testing to calculate and interpret polygenic risk scores. BioRxiv 861831.
20. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., and O'Connell, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
21. Rimfeld, K., Malanchini, M., Spargo, T., Spickernell, G., Selzam, S., McMillan, A., Dale, P.S., Eley, T.C., and Plomin, R. (2019). Twins early development study: A genetically sensitive investigation into behavioral and cognitive development from infancy to emerging adulthood. Twin Res. Hum. Genet. 1–6.
22. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., and Sharp, K. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet.
23. consortium, U. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
24. Glanville, K.P., Coleman, J.R.I., Hanscombe, K.B., Euesden, J., Choi, S.W., Purves, K.L., Breen, G., Air, T.M., Andlauer, T.F.M., Baune, B.T., et al. (2020). Classical human leukocyte antigen alleles and C4 haplotypes are not significantly associated with depression. Biol. Psychiatry *87*, 419–430.
25. Selzam, S., McAdams, T.A., Coleman, J.R.I., Carnell, S., O'Reilly, P.F., Plomin, R., and Llewellyn, C.H. (2018). Evidence for gene-environment correlation in child feeding: Links between common genetic variation for BMI in children and parental feeding practices. PLoS Genet. *14*,.
26. Selzam, S., Ritchie, S.J., Pingault, J.-B., Reynolds, C.A., O'Reilly, P.F., and Plomin, R. (2019). Comparing within-and between-family polygenic score prediction. Am. J. Hum. Genet. *105*, 351–363.
27. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 1.
28. Zeng, J., De Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., and McRae, A.F. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nat. Genet. *50*, 746.
29. Lee, S.H., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2012). A better coefficient of determination for genetic profile analysis. Genet. Epidemiol. *36*, 214–224.
30. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Statistical Methodol. *67*, 301–320.
31. Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. Psychol. Bull. *87*, 245.
32. Cooper, H., Hedges, L.V., and Valentine, J.C. (2009). The handbook of research synthesis and meta-analysis 2nd edition. In The Hand. of Res. Synthesis and Meta-Analysis, 2nd Ed., (Russell Sage Foundation), pp. 1–615.
33. Chen, C., Han, J., Hunter, D.J., Kraft, P., and Price, A.L. (2015). Explicit modeling of ancestry improves polygenic risk scores and BLUP prediction. Genet. Epidemiol. *39*, 427–438.
34. Krapohl, E., Patel, H., Newhouse, S., Curtis, C.J., von Stumm, S., Dale, P.S., Zabaneh, D., Breen, G., O'Reilly, P.F., and Plomin, R. (2018). Multi-polygenic score approach to trait

prediction. Mol. Psychiatry *23*, 1368–1374.
35. Grotzinger, A.D., Rhemtulla, M., de Vlaming, R., Ritchie, S.J., Mallard, T.T., Hill, W.D., Ip, H.F., Marioni, R.E., McIntosh, A.M., and Deary, I.J. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nat. Hum. Behav. *3*, 513–525.
36. Maier, R.M., Zhu, Z., Lee, S.H., Trzaskowski, M., Ruderfer, D.M., Stahl, E.A., Ripke, S., Wray, N.R., Yang, J., and Visscher, P.M. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nat. Commun. *9*, 1–17.
37. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., and Furlotte, N.A. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet. *50*, 229–237.
38. Marquez-Luna, C., Gazal, S., Loh, P.-R., Furlotte, N., Auton, A., Price, A.L., and Team, 23andMe Research (2018). Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. BioRxiv 375337.

Table 1. Sample size of target sample phenotypes after quality control

UKB Phenotype	Description	Total sample size	No. of cases	No. of controls
Depression	Major depression	49995	24999	24996
Intelligence	Fluid intelligence	49998	NA	NA
BMI	Body Mass Index	49993	NA	NA
Height	Height	49993	NA	NA
T2D	Type-2 Diabetes	49990	35102	14888
CAD	Coronary Artery Disease	49991	24998	24993
IBD	Inflammatory Bowel Disease	50000	46539	3461
MultiScler	Multiple Sclerosis	50000	48863	1137
RheuArth	Rheumatoid Arthritis	50000	46592	3408
TEDS Phenotype				
GCSE	Mean GCSE scores	7296	NA	NA
ADHD	ADHD symptoms	7880	NA	NA
BMI21	Body Mass Index at age 21	5220	NA	NA
Height21	Height at age 21	5455	NA	NA

Table 2. Description of polygenic scoring approaches.

Method	GWAS-effect size adjustment	Multiple tuning parameters	Pseudo-validation option	Software	Variant-level data required[A]	Description	Parameters	LD-reference
pT+clump[27]	No	Yes	No	PLINK	Direction of effect	LD-based clumping and *p*-value thresholding	10 nested *p*-value thresholds: 1e-8, 1e-6, 1e-4, 1e-2, 0.1, 0.2, 0.3, 0.4, 0.5, 1	EUR 1KG, EUR 10K UKB
lassosum[10]	Yes	Yes	Yes	lassosum	Direction of effect	Lasso regression-based	80 s and lambda combinations: s = 0.2, 0.5, 0.9, 1 lambda = exp(seq(log(0.001), log(0.1), length.out=20))[D]	EUR 1KG, EUR 10K UKB
PRScs[11]	Yes	Yes	Yes	PRScs	Direction of effect	Bayesian shrinkage	4 global shrinkage parameters (phi) = 1e-6, 1e-4, 1e-2, 1	PRScs-provided
SBLUP[9]	Yes	No	Yes (only option)	GCTA	BETA[B], SE, MAF	Best Linear Unbiased Prediction	NA	EUR 1KG, EUR 10K UKB
SBayesR[12]	Yes	No	Yes (only option)	GCTB	BETA[B], SE, MAF	Bayesian shrinkage	NA	EUR 1KG, EUR 10K UKB, GCTB-provided
LDPred[8]	Yes	Yes	Yes[C]	LDPred	Direction of effect	Bayesian shrinkage	7 non-zero effect fractions (p) = 3e-3, 1e-3, 3e-2, 1e-2, 3e-1, 1e-1, 1	EUR 1KG

Note. Default or recommended parameters were used for all methods.

[A] All methods require SNP ID, A1, A2 and P information for each variant.

[B] BETA and SE estimates were estimated using Z scores, sample size and reference sample MAF information.

[C] Infintesimal model is considered as a pseudovalidation approach.

[D] lassosum lambda values described using R code.

Table 3. Average test-set correlation between predicted and observed values across phenotypes.

Method	Group	CrossVal R (SE)	IndepVal R (SE)
pT+clump	10FCVal	0.154 (0.002)	0.153 (0.005)
pT+clump	Multi-PRS	0.173 (0.002)	0.176 (0.005)
lassosum	10FCVal	0.184 (0.002)	0.183 (0.005)
lassosum	Multi-PRS	0.193 (0.002)	0.193 (0.005)
lassosum	PseudoVal	0.154 (0.002)	0.155 (0.005)
PRScs	10FCVal	0.184 (0.002)	0.185 (0.005)
PRScs	Multi-PRS	0.19 (0.002)	0.19 (0.005)
PRScs	PseudoVal	0.184 (0.002)	0.183 (0.005)
SBLUP	PseudoVal	0.162 (0.002)	0.162 (0.005)
SBayesR	PseudoVal	0.083 (0.002)	0.087 (0.005)
LDPred	10FCVal	0.177 (0.002)	0.178 (0.005)
LDPred	Multi-PRS	0.179 (0.002)	0.18 (0.005)
LDPred	PseudoVal	0.162 (0.002)	0.161 (0.005)

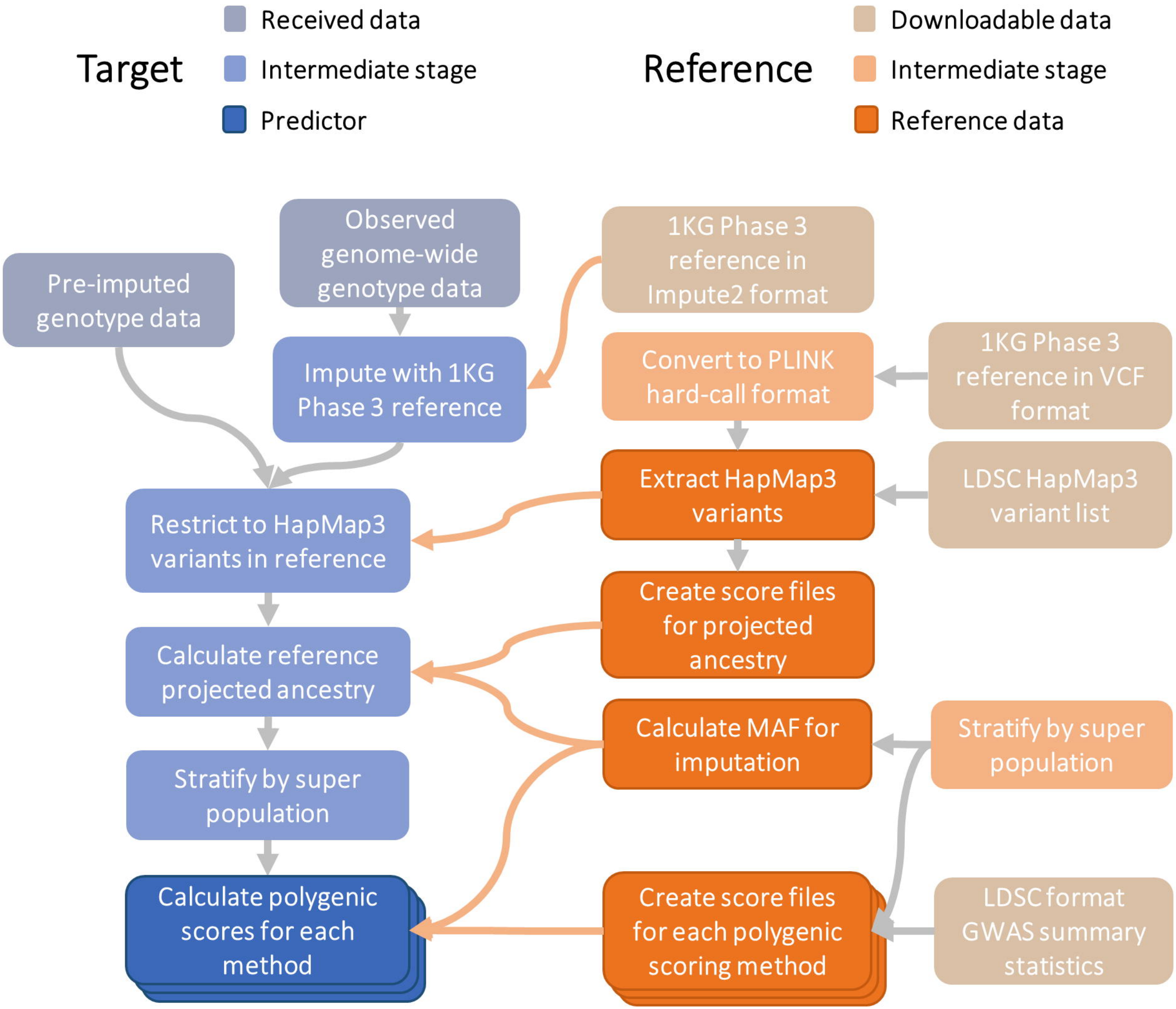
Note. This table shows results based on the UKB target sample and 1000 genomes reference. 10FCVal = Single polygenic score based on the optimal parameter as identified using 10-fold cross-validation. Multi-PRS = Elastic net model containing polygenic scores based on a range of parameters, with elastic net shrinkage parameters derived using 10-fold cross-validation. PseudoVal = Single polygenic score based on the predicted optimal parameter as identified using pseudovalidation, which requires no tuning sample.

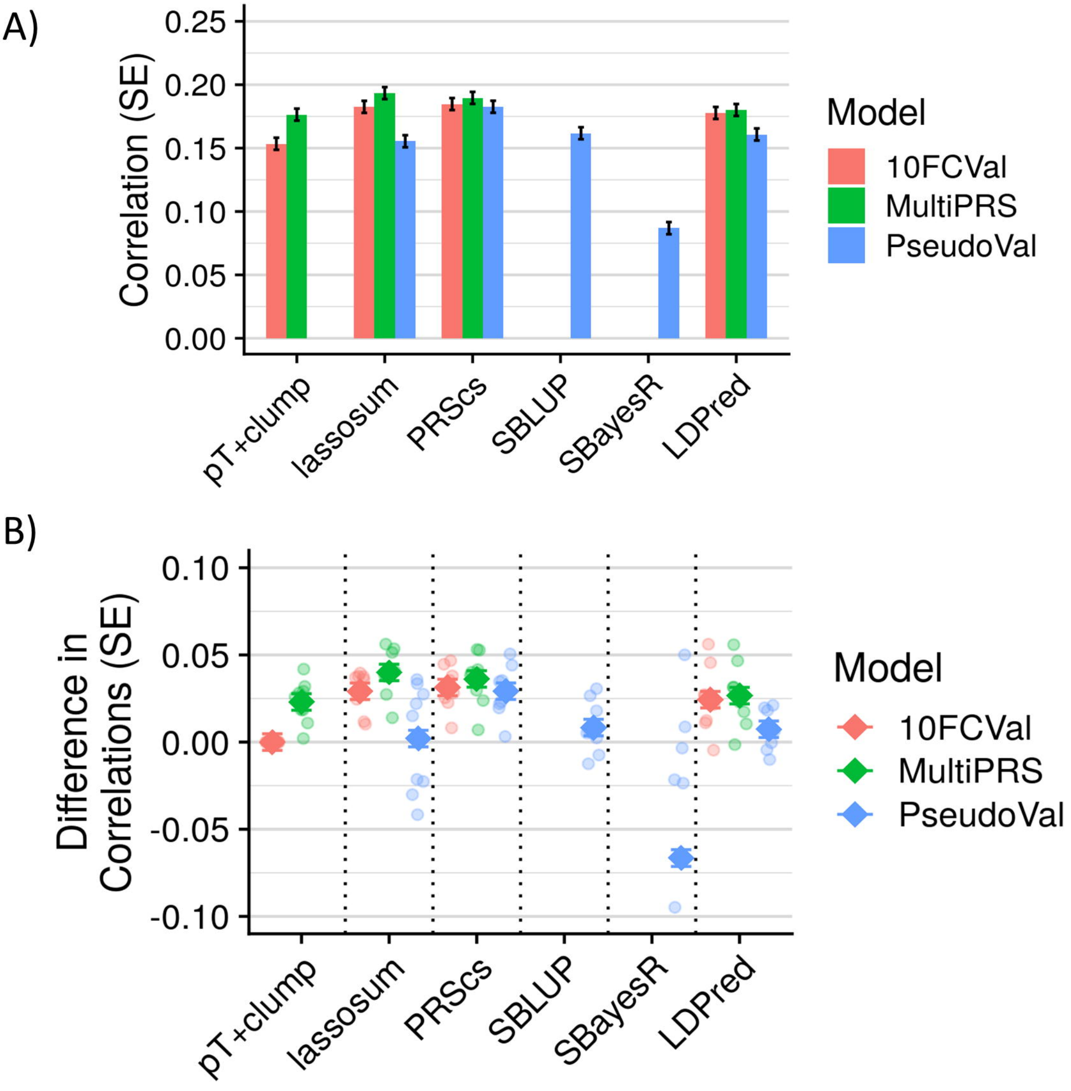
Figure Legends:

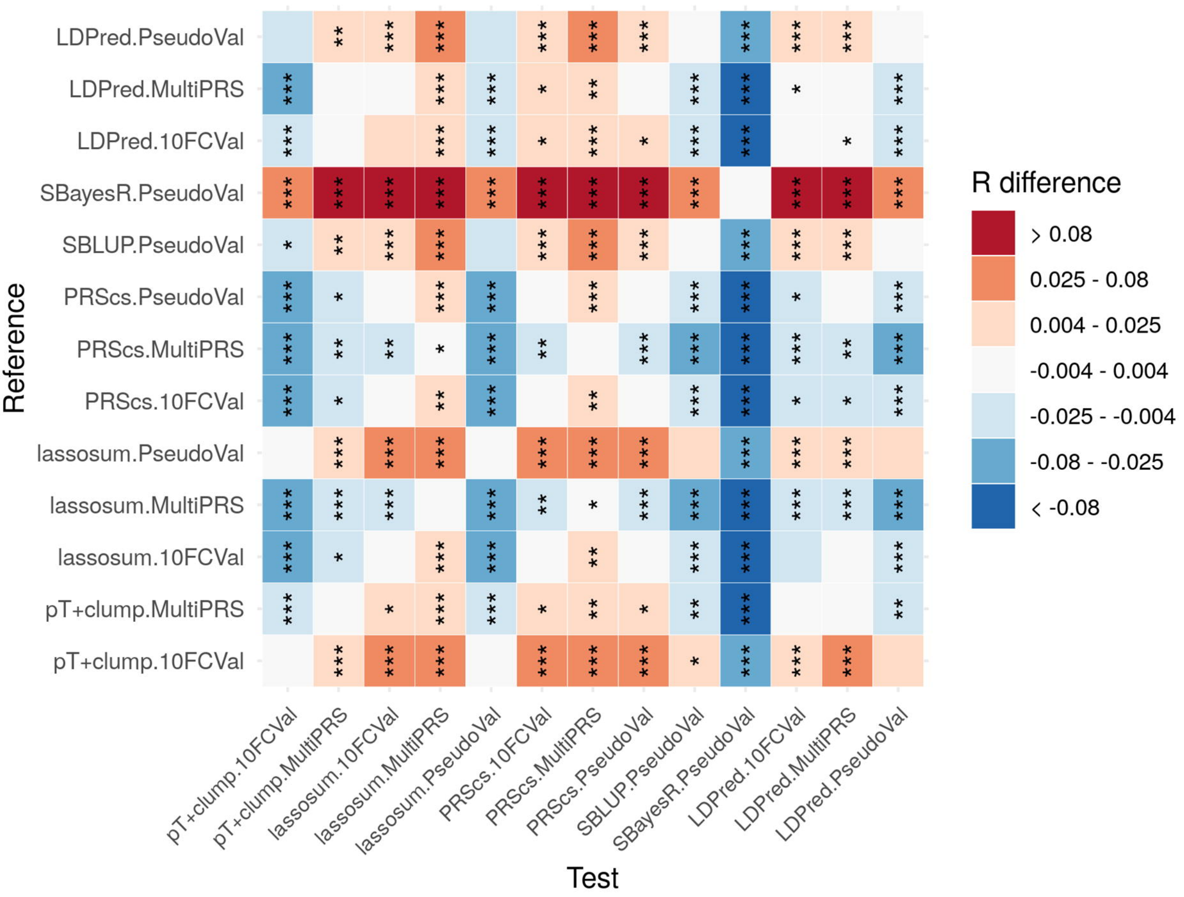
Figure 1. Schematic diagram of reference-standardized polygenic scoring. 1KG = 1000 Genomes; LDSC = Linkage Disequiibrium Score Regression; MAF = Minor allele Frequency.

Figure 2. Polygenic scoring methods comparison for UKB target sample with 1KG reference. A) Average test-set correlation between predicted and observed values across phenotypes. B) Average difference between observed-prediction correlations for the best pT+clump polygenic score and all other methods. The average difference across phenotypes are shown as diamonds and the difference for each phenotype shown as transparent circles. SBayesR phenotype-specific correlation differences < -0.1 are omitted. Shows only results based on the UKB target sample when using the 1KG reference as other results were highly concordant. Error bars indicate standard error of correlations for each method. 10FCVal represents a single polygenic score based on the optimal parameter as identified using 10-fold cross-validation. Multi-PRS represents an elastic net model containing polygenic scores based on a range of parameters, with elastic net shrinkage parameters derived using 10-fold cross-validation. PseudoVal represents a single polygenic score based on the predicted optimal parameter as identified using pseudovalidation, which requires no tuning sample.

*Figure 3. Average test-set observed-expected correlation difference between all methods with significance value. Correlation difference = Test correlation – Reference correlation. For columns, red/orange coloring indicates the Test method performed better than the Test method (horizontal). Shows only results based on the UKB target sample when using the 1KG reference as other results were highly concordant. *=$p<0.05$. **=$p<1\times10^{-3}$. ***= $p<1\times10^{-6}$. P-values are two-sided. 10FCVal represents a single polygenic score based on the optimal parameter as identified using 10-fold cross-validation. Multi-PRS represents an elastic net model containing polygenic scores based on a range of parameters, with elastic net shrinkage parameters derived using 10-fold cross-validation. PseudoVal represents a single polygenic score based on the predicted optimal parameter as identified using pseudovalidation, which requires no tuning sample.*


Target
Received data
Intermediate stage
Predictor
Reference
Downloadable data
Intermediate stage
Reference data
Pre-imputed genotype data
Observed genome-wide genotype data
Impute with 1KG Phase 3 reference
1KG Phase 3 reference in Impute2 format
Convert to PLINK hard-call format
1KG Phase 3 reference in VCF format
Extract HapMap3 variants
LDSC HapMap3 variant list
Restrict to HapMap3 variants in reference
Create score files for projected ancestry
Calculate reference projected ancestry
Calculate MAF for imputation
Stratify by super population
Stratify by super population
Calculate polygenic scores for each method
Create score files for each polygenic scoring method
LDSC format GWAS summary statistics


A)
Correlation (SE)
0.25
0.20
0.15
0.10
0.05
0.00
pT+clump
lassosum
PRScs
SBLUP
SBayesR
LDPred
Model
10FCVal
MultiPRS
PseudoVal
B)
Difference in Correlations (SE)
0.10
0.05
0.00
-0.05
-0.10
pT+clump
lassosum
PRScs
SBLUP
SBayesR
LDPred
Model
10FCVal
MultiPRS
PseudoVal


Reference
LDPred.PseudoVal
LDPred.MultiPRS
LDPred.10FCVal
SBayesR.PseudoVal
SBLUP.PseudoVal
PRScs.PseudoVal
PRScs.MultiPRS
PRScs.10FCVal
lassosum.PseudoVal
lassosum.MultiPRS
lassosum.10FCVal
pT+clump.MultiPRS
pT+clump.10FCVal
Test
pT+clump.10FCVal
pT+clump.MultiPRS
lassosum.10FCVal
lassosum.MultiPRS
lassosum.PseudoVal
PRScs.10FCVal
PRScs.MultiPRS
PRScs.PseudoVal
SBLUP.PseudoVal
SBayesR.PseudoVal
LDPred.10FCVal
LDPred.MultiPRS
LDPred.PseudoVal
R difference
> 0.08
0.025 - 0.08
0.004 - 0.025
-0.004 - 0.004
-0.025 - -0.004
-0.08 - -0.025
< -0.08