Significance tests for R^2 of out-of-sample prediction using polygenic scores

Md. Moksedul Momin,[1,2,3,4*] Soohyun Lee,[5] Naomi R Wray[6,7] and S. Hong Lee[1,2,4*]

[1] Australian Centre for Precision Health, University of South Australia Cancer Research Institute, University of South Australia, Adelaide, SA, 5000, Australia
[2] UniSA Allied Health and Human Performance, University of South Australia, Adelaide, SA, 5000, Australia
[3]Department of Genetics and Animal Breeding, Faculty of Veterinary Medicine, Chattogram Veterinary and Animal Sciences University (CVASU), Khulshi, Chattogram, 4225, Bangladesh
[4]South Australian Health and Medical Research Institute (SAHMRI), University of South Australia, Adelaide, SA, 5000, Australia
[5]Division of Animal Breeding and Genetics, National Institute of Animal Science (NIAS), South Korea
[6]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia
[7]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia

Correspondance : Cvasu.Momin@gmail.com and Hong.Lee@unisa.edu.au

Abstract

The coefficient of determination (R^2) is a well-established measure to indicate the predictive ability of polygenic scores (PGS). However, the sampling variance of R^2 is rarely considered so that 95% confidence intervals (CI) are not usually reported. Moreover, when comparisons are made between PGS based on different discovery samples, the sampling covariance of R^2 is necessary to test the difference between them. Here, we show how to estimate the variance and covariance of R^2 values to assess the 95% CI and p-value of the R^2 difference. We apply this approach to real data to predict into 28,880 European participants using UK Biobank (UKBB) and Biobank Japan (BBJ) GWAS summary statistics for cholesterol and BMI. We quantify the significantly higher predictive ability of UKBB PGS compared to BBJ PGS (p-value 7.6e-31 for cholesterol and 1.4e-50 for BMI). A joint model of UKBB and BBJ PGS significantly improves the predictive ability, compared to a model of UKBB PGS only (p-value 3.5e-05 for cholesterol and 1.3e-28 for BMI). The proposed approach can also be applied to testing a significant difference between R^2 values across different p-value thresholds. We also show that the predictive ability of regulatory SNPs is significantly enriched than non-regulatory SNPs for cholesterol (p-value 2.6e-19 for UKBB and 8.7e-08 for BBJ). We suggest that the proposed approach (available in R package 'r2redux') should be used to test the statistical

significance of difference between pairs of PGS, which may help to draw a correct conclusion about the predictive ability of PGS.

Introduction

Complex traits are affected by many risk factors including polygenic effects [1-3]. Genetic profile analysis can quantify how polygenic effects are associated with future disease risk at the individual and population levels[4; 5]. Genetic profiling has potential benefits that can help people make informed decisions when they manage their health and medical care [6-8].

Genome-wide association studies (GWAS) have provided an opportunity to estimate genetic profile or polygenic scores (PGS) that can make individual risk predictions from genetic data [4; 9-14]. Typically, the effects of genome-wide single nucleotide polymorphisms (SNPs) associated with complex traits are estimated in a discovery dataset, which are projected in an independent target dataset. Then, for each individual in the target samples the weighted genotypic coefficients according to the projected SNP effects (i.e. PGS) are derived and correlated with outcome (trait including affected/unaffected for disease) to quantify the prediction accuracy. The squared correlation or coefficient of determination (R^2) have been a useful measure to quantify the reliability of PGS. Note that R^2 is equivalent to squared regression coefficients if the dependent and explanatory variables are column-standardised [15].

Previously, we introduced a novel measure of R^2 on the liability scale that can be comparable across different models and scales [16] when using disease traits or ascertained case-control data. Choi et al. [12] reported that this R^2 measure on the liability scale outperforms the widely used Nagelkerke pseudo R^2 in controlling for bias due to ascertained case-control samples. Nagelkerke pseudo R^2 is not independent of the proportion of cases in the sample. In contrast, R^2 on the liability scale is not dependent on the proportion of cases in the sample, but does require an estimate of the lifetime population prevalence of diseases.

Wand et al. [11] suggested that any PGS study should report R^2 as an indicator of the predictive ability. Choi et al. [12] concluded that R^2 is a useful metric to measure association and goodness of fit in the interpretation of PGS predictions. Many studies have demonstrated the predictive ability of PGS, using R^2 [12; 13; 17; 18]. However, the variance of R^2 [15] has been rarely studied especially in the PGS analyses although it is the crucial parameter to estimate confidence

intervals of R^2. Furthermore, estimates of the covariance between a pair of R^2 values (e.g., from two sets of PGS) are needed to assess if they are significantly different to each other, or if the ratio of two R^2 values is significantly deviated from the expectation. This significance test for the difference or ratio is important when comparing two or multiple sets of PGS that are derived from different sets of SNPs, e.g., genomic partitioning or genome-wide association p-value thresholds (p_T) analysis.

In this study, we use R^2 measures and their variance-covariance matrix to assess if the predictive abilities of PGS based on different sources are significantly different to each other. We derive the variance and covariance of R^2 values to generate estimates of its 95% CI and p-value of the R^2 difference, considering two sets of dependent or independent PGS. We also derive the variance and covariance (i.e. information matrix) of squared regression coefficients in a multiple regression model, testing if the proportion of the squared regression coefficient attributable to SNPs in the regulatory region is significantly higher than expected (i.e. PGS-based genomic partitioning method). We apply this approach to real data to predict 28,880 European individuals using UK Biobank (UKBB) and Biobank Japan (BBJ) GWAS summary statistics for cholesterol and BMI.

Methods

PGS models

We use a linear model that regresses the observed phenotypes on a single or multiple sets of PGS. It is assumed that the phenotypes are already adjusted for other non-genetic and environmental factors (e.g. demographic variables, ancestry principal components (PCs)).

A PGS model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{ß} + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is the vector of standardised phenotypes of trait, $\mathbf{X}$ is a column-standardised N x M matrix including M sets of PGS, $\boldsymbol{ß}$ is the vector of regression coefficients and $\mathbf{e}$ is the vector of residuals. For example, with two sets of PGS (M=2), $\mathbf{X}$ and $\widehat{\boldsymbol{\beta}}$ can be expressed as

$$\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}],$$

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \quad (2)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} (\boldsymbol{\Sigma_{11}}) & (\boldsymbol{\Sigma_{12}}) \\ (\boldsymbol{\Sigma_{21}}) & (\boldsymbol{\Sigma_{22}}) \end{bmatrix} = \begin{bmatrix} (\mathbf{1}) & (r_{y,x_1} & r_{y,x_2}) \\ \begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \end{pmatrix} & \begin{pmatrix} \mathbf{1} & r_{x_1,x_2} \\ r_{x_1,x_2} & \mathbf{1} \end{pmatrix} \end{bmatrix} \quad (3)$$

where r_{y,x_1}, r_{y,x_2} and r_{x_1,x_2} are correlations between **y** and the first PGS ($\mathbf{x_1}$), **y** and the second PGS ($\mathbf{x_2}$), and between the two PGS ($\mathbf{x_1}$ and $\mathbf{x_2}$), respectively, in the sample. Using $\widehat{\boldsymbol{\beta}}$ that are estimated in the multiple regression (eq. 2), the expected phenotypes ($\hat{\mathbf{y}}$) can be obtained as

$$\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}.$$

The coefficient of determination for this multiple regression model with $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}]$ in eq. (1) can be written as

$$r^2_{y,(x_1,x_2)} = 1 - \mathrm{E}(\mathbf{y} - \hat{\mathbf{y}})^2 = \mathrm{E}(\hat{\mathbf{y}}^2) = \hat{\beta}_1^2 + \hat{\beta}_2^2 + 2r_{x_1,x_2}\hat{\beta}_1\hat{\beta}_2. \quad (4)$$

With a single set of PGS, i.e. M=1 and $\mathbf{X} = [\mathbf{x_1}]$ or $[\mathbf{x_2}]$ in eq. (1), the expression of R^2 can be reduced as

$r^2_{y,x_1} = \mathrm{E}(\hat{\mathbf{y}}^2) = \hat{\beta}_1^2$ with $\mathbf{X} = [\mathbf{x_1}]$

or

$r^2_{y,x_2} = \mathrm{E}(\hat{\mathbf{y}}^2) = \hat{\beta}_2^2$ with $\mathbf{X} = [\mathbf{x_2}]$.

Variance of R^2

The distribution of R^2 can be transformed to a non-central χ^2 distribution with mean = M+ λ and variance = $2 \times (M + 2\lambda)$ where $\lambda = \frac{N \times R^2}{1-R^2}$ is the non-centrality parameter. For example, the variance of the transformed value for r^2_{y,x_1} is

$$var\left[\left(\frac{\hat{\beta}_1}{sd(\hat{\beta}_1)}\right)^2\right] = \frac{1}{var(\hat{\beta}_1)^2} var\left(\hat{\beta}_1^2\right) = 2(M + 2\lambda).$$

Therefore, $var\left(r^2_{y,x_1}\right) = var\left(\hat{\beta}_1^2\right) = 2var\left(\hat{\beta}_1\right)^2(M + 2\lambda)$ (5)

where $var\left(\hat{\beta}_1\right) = 1/N \cdot (1 - \rho^2_{y,x_1})^2$ and ρ^2_{y,x_1} is the squared correlation in the population and can be approximated as $\rho^2_{y,x_1} \approx r^2_{y,x_1}$ [19; 20].

In a similar manner, eq. (5) can be extended to multiple explanatory variables as

$$var\left(r^2_{y,(x_1,x_2,\ldots,x_M)}\right) \approx 2\left[\frac{1}{N}\cdot\left(1-r^2_{y,(x_1,x_2,\ldots,x_M)}\right)^2\right]^2 (M+2\lambda) \quad (6)$$

Wishart et al. [21] introduced a formula to obtain the variance of R^2 as

$$Var(R^2) = \frac{[4\times R^2\times(1-R^2)^2\ \times\{N-(M+1)\}^2]}{[(N^2-1)\times(N+3)]}$$

which provides an equivalent estimate as in eq. (6). The s.e. of R^2 estimate is the square root of $var(R^2)$.

Variance of the difference between two R^2 values

Following Olkin and Finn [15], we use the delta method to estimate the variance of the difference between R^2 values based on two sets of PGS ($\mathbf{x_1}$ and $\mathbf{x_2}$). Assuming that the difference of R^2 values can be formulated as a function of the correlations, i.e. $f\left(r_{y,x_1}, r_{y,x_2}, r_{x_1,x_2}\right)$, the delta method approximates the variance of the difference as

$$var(f) = \boldsymbol{\theta}'\boldsymbol{\Omega}\boldsymbol{\theta} \quad (7)$$

where $\boldsymbol{\theta}' = \left(\frac{\partial f}{\partial r_{y,x_1}}, \frac{\partial f}{\partial r_{y,x_2}}, \frac{\partial f}{\partial r_{x_1,x_2}}\right)$ is the derivatives of *f* with respect to the correlations (8)

and

$$\boldsymbol{\Omega} = \begin{bmatrix} var(r_{y,x_1}) & cov(r_{y,x_1}, r_{y,x_2}) & cov(r_{y,x_1}, r_{x_1,x_2}) \\ cov(r_{y,x_1}, r_{y,x_2}) & var(r_{y,x_2}) & cov(r_{y,x_2}, r_{x_1,x_2}) \\ cov(r_{y,x_1}, r_{x_1,x_2}) & cov(r_{y,x_2}, r_{x_1,x_2}) & var(r_{x_1,x_2}) \end{bmatrix}.$$

Each element of $\boldsymbol{\Omega}$ is shown in Olkin and Finn [15] (also see Appendix).

From eq. (7), the following variances of differences can be estimated and used in our PGS analyses.

1. **R^2 difference when using different discovery samples to generate the PGS**

$var\left(r^2_{y,x_1} - r^2_{y,x_2}\right)$ with $f\left(r_{y,x_1}, r_{y,x_2}, r_{x_1,x_2}\right) = r^2_{y,x_1} - r^2_{y,x_2}$, (9)

which allows us to compare two PGS models that are not nested to each other (see 'R^2 difference when using different information sources' in Results section), for which the conventional log-likelihood ratio test cannot be applied.

In eq. (9), the values of $r^2_{y,x_1} - r^2_{y,x_2}$ from random samples in the population are normally distributed when the sample size is sufficient [15]. Assuming that our PGS analysis is sufficiently powered (N > 25,000), the p-value for the significance test of the difference can be derived from

$$\frac{(r^2_{y,x_1} - r^2_{y,x_2})^2}{var(r^2_{y,x_1} - r^2_{y,x_2})} \sim \chi^2_1$$

and the 95% confidence interval is

$$\left[(r^2_{y,x_1} - r^2_{y,x_2}) - 1.96\sqrt{var(r^2_{y,x_1} - r^2_{y,x_2})},\quad (r^2_{y,x_1} - r^2_{y,x_2}) + 1.96\sqrt{var(r^2_{y,x_1} - r^2_{y,x_2})}\right] \quad (10)$$

2. **R^2 difference when using nested models**

$var(r^2_{y,(x_1,x_2)} - r^2_{y,x_2})$ with $f(r_{y,x_1}, r_{y,x_2}, r_{x_1,x_2}) = r^2_{y,(x_1,x_2)} - r^2_{y,x_2} = \hat{\beta}_1^2 + \hat{\beta}_2^2 + 2r_{x_1,x_2}\hat{\beta}_1\hat{\beta}_2 - r^2_{y,x_2}$ (11)

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimated regression coefficients from a multiple regression (eq. 2), calculated from $\mathbf{\Sigma}$ (see eq. 2–4). Again, the derivative with respect to each of the correlations can be obtained for this function (eq. 8). Note that the comparison between $r^2_{y,(x_1,x_2)}$ and r^2_{y,x_2} is equivalent to the log-likelihood ratio test (i.e. $\mathbf{y} = \mathbf{x_1}\beta_1 + \mathbf{x_2}\beta_2 + \mathbf{e}$ vs. $\mathbf{y} = \mathbf{x_2}\beta_2 + \mathbf{e}$)[15].

The values of $r^2_{y,(x_1,x_2)} - r^2_{y,x_2}$ in eq. (11) from random samples in the population follows a non-central chi-squared distribution with a non-centrality parameter $\lambda = N \times \frac{r^2_{y,(x_1,x_2)} - r^2_{y,x_2}}{1-(r^2_{y,(x_1,x_2)} - r^2_{y,x_2})^2}$. The p-value for the significance test of the difference can be derived from

$$\lambda \sim \chi^2_1$$

and the 95% confidence interval is

$$\left[(r^2_{y,(x_1,x_2)} - r^2_{y,x_2}) + \sqrt{var(r^2_{y,(x_1,x_2)} - r^2_{y,x_2})}\frac{\xi_{97.5\%} - \lambda - 1}{\sqrt{2(1+2\lambda)}},\quad (r^2_{y,(x_1,x_2)} - r^2_{y,x_2}) + \sqrt{var(r^2_{y,(x_1,x_2)} - r^2_{y,x_2})}\frac{\xi_{2.5\%} - \lambda - 1}{\sqrt{2(1+2\lambda)}}\right] \quad (12)$$

where $\xi_{\%}$ is the value at the percentile of the inverse of non-central chi-squared cumulative distribution function with mean = $\lambda + 1$ and d.f. = 1.

When the sample size is large, the values of $r^2_{y,(x_1,x_2)} - r^2_{y,x_2}$ from random samples in the population are normally distributed [15]. The p-value for the significance test of the difference can also be derived from

$$\frac{(r^2_{y,(x_1,x_2)} - r^2_{y,x_2})^2}{var\left(r^2_{y,(x_1,x_2)} - r^2_{y,x_2}\right)} \sim \chi^2_1$$

and the 95% confidence interval is

$$\left[\left(r^2_{y,(x_1,x_2)} - r^2_{y,x_2}\right) - 1.96\sqrt{var\left(r^2_{y,(x_1,x_2)} - r^2_{y,x_2}\right)},\quad \left(r^2_{y,(x_1,x_2)} - r^2_{y,x_2}\right) + 1.96\sqrt{var\left(r^2_{y,(x_1,x_2)} - r^2_{y,x_2}\right)}\right] \quad (13)$$

Note that eq. 12 and 13 are equivalent when the sample size is sufficient [15].

3. R^2 difference when using two independent sets of PGS

In this case, there is no correlation structure between two independent sets of PGS ($r_{x_1,x_2} = 0$, e.g., male and female PGS), therefore, the variance of R^2 difference is simply the sum of the variances of each R^2 value, which can be obtained from eq. (5). For example, assuming $r_{x_1,x_2} = 0$, the variance of R^2 difference can be written as

$$var\left(r^2_{y_1,x_1} - r^2_{y_2,x_2}\right) = 2\left[\frac{1}{N_1} \cdot \left(1 - r^2_{y_1,(x_1)}\right)^2\right]^2 (1 + 2\lambda_1) + 2\left[\frac{1}{N_2} \cdot \left(1 - r^2_{y_2,(x_1)}\right)^2\right]^2 (1 + 2\lambda_2) \quad (14)$$

where y_1 and y_2 are the vectors of standardised phenotypes and N_1 and N_2 are the sample sizes for the two independent sets of PGS. The non-centrality parameters (λ_1 and λ_2) for two dependent PGS can be written as

$$\lambda_1 = \frac{N_1 \times r^2_{y_1,x_1}}{1 - r^2_{y_1,x_1}} \text{ and } \lambda_2 = \frac{N_2 \times r^2_{y_2,x_2}}{1 - r^2_{y_2,x_2}}$$

p-value for the significance test of the difference can be derived from $\frac{(r^2_{y_1,x_1} - r^2_{y_2,x_2})^2}{var(r^2_{y_1,x_1} - r^2_{y_2,x_2})} \sim \chi^2_1$

and the 95% confidence interval is

$$\left[\left(r^2_{y_1,x_1} - r^2_{y_2,x_2}\right) - 1.96\sqrt{var\left(r^2_{y_1,x_1} - r^2_{y_2,x_2}\right)},\ \left(r^2_{y_1,x_1} - r^2_{y_2,x_2}\right) + 1.96\sqrt{var\left(r^2_{y_1,x_1} - r^2_{y_2,x_2}\right)}\right]$$

(15)

4. PGS-based genomic partitioning analysis

$var(\hat{\beta}_1^2 - \hat{\beta}_2^2)$ with $f(r_{y,x_1}, r_{y,x_2}, r_{x_1,x_2}) = \hat{\beta}_1^2 - \hat{\beta}_2^2$ (16)

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimated regression coefficients from a multiple regression (eq. 2), calculated from **Σ** (eq. 3). Therefore, it is possible to get the derivative with respect to each of the correlations, r_{y,x_1}, r_{y,x_2} and r_{x_1,x_2} in eq. (8). $var(\hat{\beta}_1^2)$ and $var(\hat{\beta}_2^2)$ can be also obtained in a similar manner. Thus, we can get the variance covariance matrix ($\mathcal{H}$), i.e. the information matrix, as

$$\mathcal{H} = \begin{bmatrix} var(\hat{\beta}_1^2) & -\frac{var(\hat{\beta}_1^2-\hat{\beta}_2^2)-var(\hat{\beta}_1^2)-var(\hat{\beta}_2^2)}{2} \\ -\frac{var(\hat{\beta}_1^2-\hat{\beta}_2^2)-var(\hat{\beta}_1^2)-var(\hat{\beta}_2^2)}{2} & var(\hat{\beta}_2^2) \end{bmatrix}.$$

The $\mathcal{H}$ can be used to estimate the variance of the difference between $\frac{\hat{\beta}_1^2}{p_{exp}}$ and $\frac{\hat{\beta}_2^2}{(1-p_{exp})}$ as

$$var\left[\frac{\hat{\beta}_1^2}{p_{exp}} - \frac{\hat{\beta}_2^2}{(1-p_{exp})}\right] = \left[\frac{var(\hat{\beta}_1^2)}{(p_{exp})^2} + \frac{var(\hat{\beta}_2^2)}{(1-p_{exp})^2} + \frac{var(\hat{\beta}_1^2-\hat{\beta}_2^2)-var(\hat{\beta}_1^2)-var(\hat{\beta}_2^2)}{p_{exp}(1-p_{exp})}\right] \quad (17)$$

where the expected proportion of phenotypic variance explained by x_1 (PGS1) can be calculated from prior information, referred to as p_{exp} = # SNPs used for PGS1 / total # SNPs. This variance can be used to test if two squared regression coefficients scaled by their expectations are significantly different to each other.

Analogous to eq. (9), the values of $\frac{\hat{\beta}_1^2}{p_{exp}} - \frac{\hat{\beta}_2^2}{(1-p_{exp})}$ in eq. (17) with random samples in the population are asymptotically normal [15]. Using a Wald test, the p-value for the significance test of the difference can be derived from

$$\frac{\left[\left(\frac{\hat{\beta}_1^2}{p_{exp}} - \frac{\hat{\beta}_2^2}{(1-p_{exp})}\right)\right]^2}{var\left(\frac{\hat{\beta}_1^2}{p_{exp}} - \frac{\hat{\beta}_2^2}{(1-p_{exp})}\right)} \sim \chi_1^2 .$$

The 95% confidence interval of the ratio is

$$\left[\left(\frac{\hat{\beta}_1^2}{p_{exp}}-\frac{\hat{\beta}_2^2}{(1-p_{exp})}\right)-1.96\sqrt{var\left(\frac{\hat{\beta}_1^2}{p_{exp}}-\frac{\hat{\beta}_2^2}{(1-p_{exp})}\right)},\ \left(\frac{\hat{\beta}_1^2}{p_{exp}}-\frac{\hat{\beta}_2^2}{(1-p_{exp})}\right)+1.96\sqrt{var\left(\frac{\hat{\beta}_1^2}{p_{exp}}-\frac{\hat{\beta}_2^2}{(1-p_{exp})}\right)}\right] \quad (18)$$

Data

The UK Biobank is a large-scale biomedical database, that comprises 0.5 million individuals who had been recruited between 2006 to 2010 and their age ranged between 40 and 69 years [22; 23]. The data consists of health-related information for samples who are genotyped for genome-wide SNPs. A stringent quality control (QC) process was applied to UKBB data that excludes individuals with non-white British ancestries, mismatched gender between reported and genotypic information, genotype call rate <0.95 and putative sex chromosome aneuploidy. The SNP QC criteria filtered out SNPs with an imputation reliability <0.6, missingness >0.05, minor allele frequency (MAF) <0.01, Hardy-Weinberg equilibrium p-value < 10^{-07}. We also applied a relatedness cut-off QC (>0.05) so that there was no high pairwise relatedness among samples. After QC, 288,792 individuals and 7,701,772 SNPs were remained.

Discovery GWAS data

Ninety percent of the individuals from the QCed data (N=288,792) were randomly selected as discovery samples (N=259,912, SNP # =7,701,772) to generate GWAS summary statistics (UKBB hereafter). For the GWAS with the 259,912 UKBB discovery samples, we used BMI and cholesterol that were adjusted for age, sex, birth year, Townsend deprivation Index (TDI), education, genotype measurement batch, assessment centre and the first 10 ancestry principal components (PCs) using a linear regression.

We also have access to Japanese Biobank (BBJ) (http://jenger.riken.jp/en/result) GWAS summary statistics (BBJ hereafter) for BMI [24] (N=158,284) and cholesterol [25] (N=128,305) for 5,961,601 SNPs.

Target data

Ten percent of the individuals from the QCed data (N=288,792) were randomly selected as an independent target dataset (N=28,880 and SNP # =7,701,772) that were non-overlapping and unrelated with the UKBB and BBJ discovery samples. In the PGS analyses, we used only

4,113,630 SNPs that were common between UKBB and BBJ GWAS data after excluding ambiguous SNPs and SNPs with any strand issue.

In the target dataset (N=28,880), the phenotypes of each trait were adjusted for age, sex, birth year, TDI, education, genotype batch, assessment centre and the first 10 PCs using a linear regression. The pre-adjusted phenotypes were correlated with PGS estimated in the following step. For each trait, we used the UKBB and BBJ GWAS summary statistics to estimate two sets of PGS (UKBB PGS vs. BBJ PGS for the target individuals (n=28,880), using PLINK2 --score function [26]. Then, we estimated the correlation between the PGS and pre-adjusted phenotypes to obtain R^2 values in the PGS analyses.

Functional annotation of the genome

We annotated the genome using pre-defined functional categories (regulatory vs. non-regulatory genomic regions) [27]. Regulatory region includes SNPs from coding regions, untranslated regions (UTR) and promotors. Non-regulatory region includes all the other regions except the regulatory region. The number of SNPs belong to regulatory and non-regulatory is 158,653 and 3,954,947 (i.e. 4% of the total SNPs are located in the regulatory region).

Simulation of dependent and explanatory variables

For a quantitative trait, we simulated dependent variable (y) and PGS (x_1 and x_2), varying the correlation structure of $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix}$ and the sample size (detailed simulation parameters are shown in Supplementary Figures 1-9). For a disease trait, the same simulation procedure was used, and the simulated quantitative phenotypes were transformed to binary responses using a liability threshold model with a population prevalence of k=0.05. For example, case-control status was assigned to individuals according to their standardised quantitative phenotypes (i.e. liability), i.e. cases have liability greater than a threshold such that the proportion of cases is k=0.05. The empirical variances of r^2_{y,x_1}, $r^2_{y,x_1} - r^2_{y,x_2}$, $r^2_{y,(x_1,x_2)} - r^2_{y,x_2}$ and $\frac{\widehat{\beta}_1^2}{p_{exp}} - \frac{\widehat{\beta}_2^2}{(1-p_{exp})}$ were obtained over 10,000 replicates, which were compared to the theoretical variances estimated using eqs. (6), (9), (11) and (17), respectively.

Results

Simulation verification

The theory of the proposed method has been explicitly verified using simulations, varying sample size and values of r^2_{y,x_1}, $r^2_{y,x_1} - r^2_{y,x_2}$, $r^2_{y,(x_1,x_2)} - r^2_{y,x_2}$ and $\frac{\hat{\beta}_1^2}{p_{exp}} - \frac{\hat{\beta}_2^2}{(1-p_{exp})}$ (**Supplementary Figures 1 – 9).** The empirical variances obtained from 10,000 simulated replicates are almost perfectly correlated with the theoretical variance for the values of r^2_{y,x_1}, $r^2_{y,x_1} - r^2_{y,x_2}$, $r^2_{y,(x_1,x_2)} - r^2_{y,x_2}$ and $\frac{\hat{\beta}_1^2}{p_{exp}} - \frac{\hat{\beta}_2^2}{(1-p_{exp})}$ when varying the sample size (**Supplementary Figures 1 – 4**) and when varying R^2 values (**Supplementary Figures 5 – 8**). When considering two independent PGS, the theoretical and empirical variances are also agreed well (**Supplementary Figure 9**).

R^2 difference when using different information sources (UKBB vs. BBJ)

It is of interest to determine whether different information sources (e.g., ancestries) have significantly different predictive abilities in PGS analyses, which can be assessed using eqs. (9) and (10). **Figure 1** illustrates that when predicting the 28,880 European target samples, the coefficient of determinations (R^2) with the UKBB and BBJ PGS were 0.024 (95% CI = 0.021-0.028) and 0.003 (95% CI = 0.002 - 0.004), respectively, for cholesterol. However, these R^2 values and CIs cannot be used to assess their difference because the two sets of PGS are not independent. Furthermore, the two PGS models with UKBB and BBJ are not nested to each other, therefore, the likelihood ratio test could not be used either. For this problem, we used eqs. (9) and (10) to obtain the variance, 95% CI (0.0247 - 0.0175) and p-value (7.6e-31) of the R^2 difference, accounting for the dependency between UKBB and BBJ PGS, for cholesterol (**Figure 1**). Similarly, the test statistics of the R^2 difference was obtained for BMI, 0.035 - 0.046 for 95% CI and p-value = 1.4e-50 (**Figure 1**).

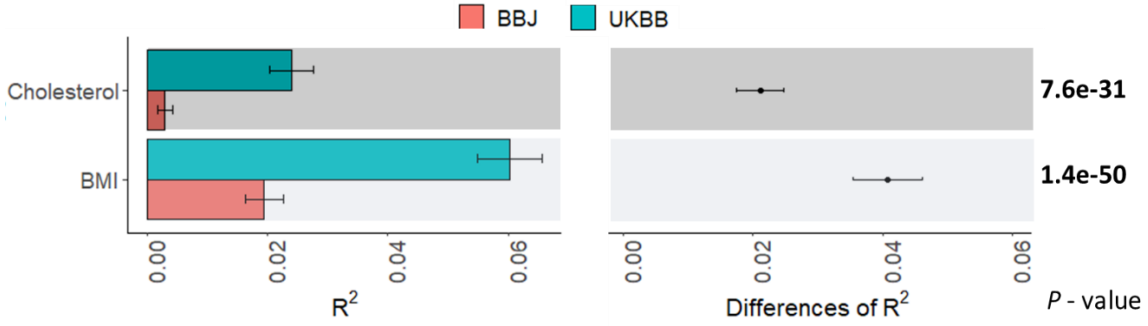

Figure 1: The predictive ability (R^2) of PGS when predicting 28,880 European individuals using UKBB or BBJ discovery GWAS dataset.
Left panel: The main bars represent R^2 values and error bars correspond 95% confidence intervals. Two sets of GWAS summary statistics were obtained from UKBB and BBJ discovery GWAS datasets to estimate two sets of PGS.
Right panel: Dot points represent the differences of R^2 values between UKBB and BBJ PGS models, and error bars indicate 95% confidence intervals of the difference.

It is also interesting to test if BBJ PGS provides a significant improvement in the predictive ability, in addition to UKBB PGS, when predicting the 28,880 European target samples. **Figure 2** compares R^2 value with each UKBB or BBJ PGS to R^2 value from a joint model fitting UKBB and BBJ PGS simultaneously. Using eq. (11) and (12), we acquired the variance, 95% CI (0.0001–0.001) and p-value (3.5e-05) of R^2 difference when comparing the joint model with a single model with UKBB, indicating that BBJ PGS contributed to a significant improvement for cholesterol. Similarly, BBJ PGS improved the predictive ability significantly (p-value = 1.3e-28) for BMI. As expected, excluding UKBB PGS from the joint model substantially decreased the prediction accuracy (p-value = 1.6e-136 for cholesterol and 3.0e-308 for BMI).

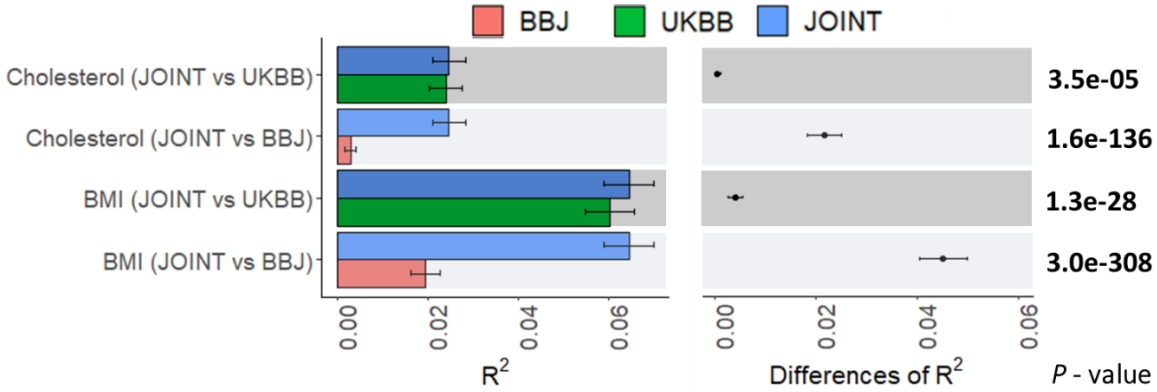

Figure 2: The predictive ability (R^2) of the UKBB or BBJ PGS model or a joint model of UKBB and BBJ when predicting 28,880 European individuals.
Left panel: The main bars represent R^2 values and error bars correspond 95% confidence intervals. Two sets of GWAS summary statistics were obtained from UKBB and BBJ discovery GWAS datasets to estimate two sets of PGS, i.e. UKBB and BBJ PGS. In addition, a joint model fitting both UKBB and BBJ PGS was compared.
Right Panel: Dot points represent the differences of R^2 values between the joint model and UKBB or BBJ PGS model, and error bars indicate 95% confidence intervals of the difference.

R^2 difference when using two independent sets of PGS (male vs. female)

We were also interested in testing if the PGS could predict the adjusted phenotypes of the target individuals equally well for males and females. In this case, there is no correlation structure between male and female PGS, therefore, the variance of R^2 difference is simply the sum of

the variances of each R^2 value, which can be obtained from eq. (5) or (6**). Supplementary Figure 10** shows that there was no significant difference between male and female PGS in their predictive ability for cholesterol and BMI whether using UKBB or BBJ discovery GWAS dataset.

PGS with genome-wide association p-value thresholds (p_T)

PGS also has been widely used to determine which p_T provides the highest prediction accuracy, for example, using PGS software such as PLINK [26; 28]. However, there is a lack of test statistics that can assess if the predictive ability of the best-performed p_T is significantly different from the other p_T. **Figure 3a** illustrates that R^2 value is the highest at $p_T = 0.3$ when predicting 28,880 European individuals in the target dataset, using BBJ discovery GWAS dataset (BMI). However, it is not clear if the predictive ability at $p_T = 0.3$ is significantly higher than the adjacent p_T (e.g. p_T = 0.2 or 0.4), and it may be important to report p_T of which the predictive ability is not statistically different from the best-performed p_T. Using eqs. (9) and (10), we assessed the significance of difference between the best-found p_T and each of the other p_T (**Figure 3b**). From this analysis, we found that the best-performed p_T was not significantly different from p_T ranging between 0.1 – 1, but significantly different from $p_T \leq 0.05$ (**Figure 3b**). When using UKBB discovery GWAS dataset to predict the 28,880 European individuals, the highest R^2 value at the p_T of 1 was significantly different from all the other p_T (**Supplementary Figure 11b).**

Interestingly, the highest R^2 value was found at p_T = 1e-04 (**Figure 4a**) when predicting the European target samples using BBJ discovery GWAS dataset for cholesterol, which was not statistically different from p_T = 0.001, but was significantly higher than the other p_T (**Figure 4b**). For the same target samples and trait, the best R^2 value was obtained from p_T = 0.01 when using UKBB discovery GWAS dataset (**Supplementary Figure 12a**). Except for p_T = 0.01, 0.05 and 0.1, R^2 values at the other p_T were significantly different from the best R^2 values (**Supplementary Figure 12b**).

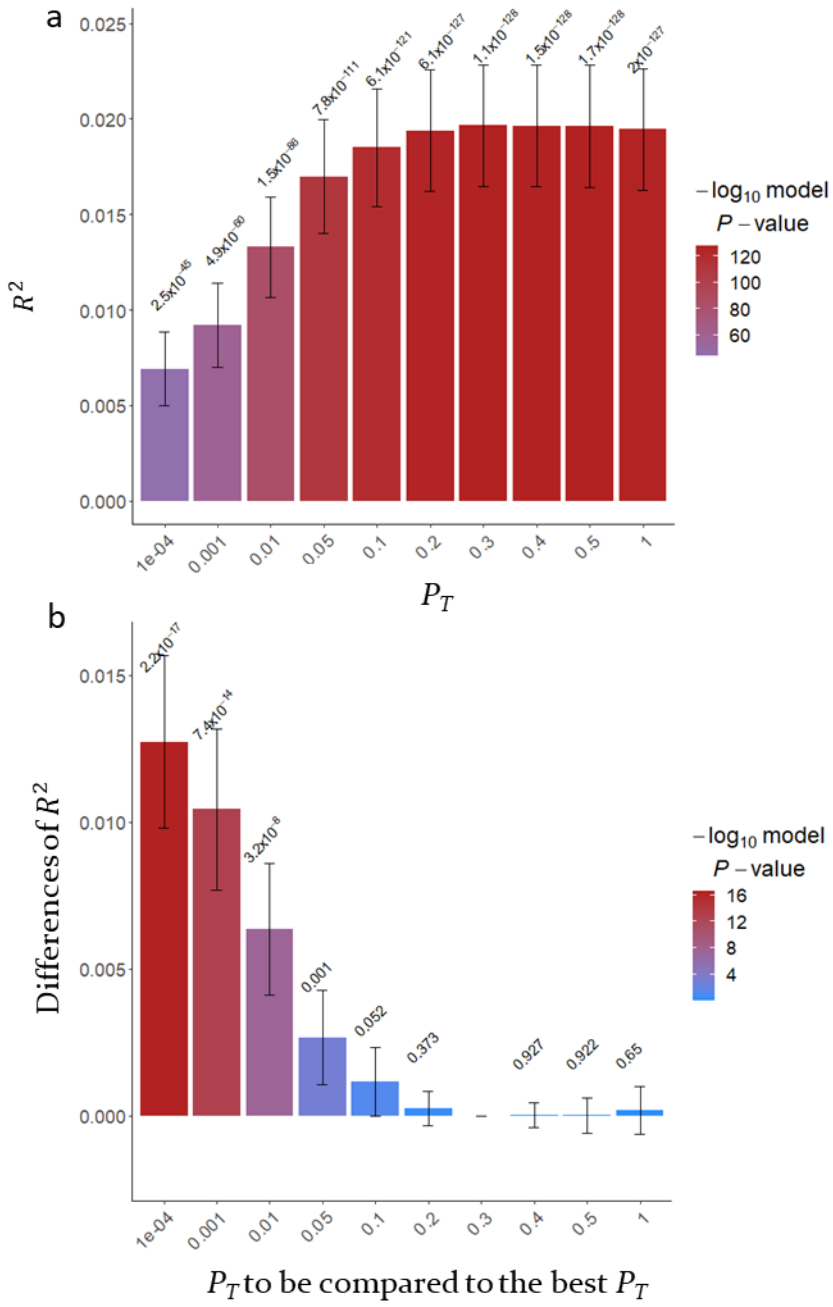

Figure 3: The predictive ability (R^2) of PGS estimated based on SNPs below p_T when predicting BMI in 28,880 European samples using BBJ discovery samples (GWAS summary statistics).

a) The main bars represent R^2 values and error bars correspond 95% confidence intervals. The values above 95% CIs are p-values indicating that R^2 values are not different from zero.

b) The main bars represent the difference of R^2 values between the corresponding p_T and the best-performed p_T and error bars indicate 95% confidence intervals. The values above 95% CIs are p-values indicating the significances of differences between the pairs of R^2 values.

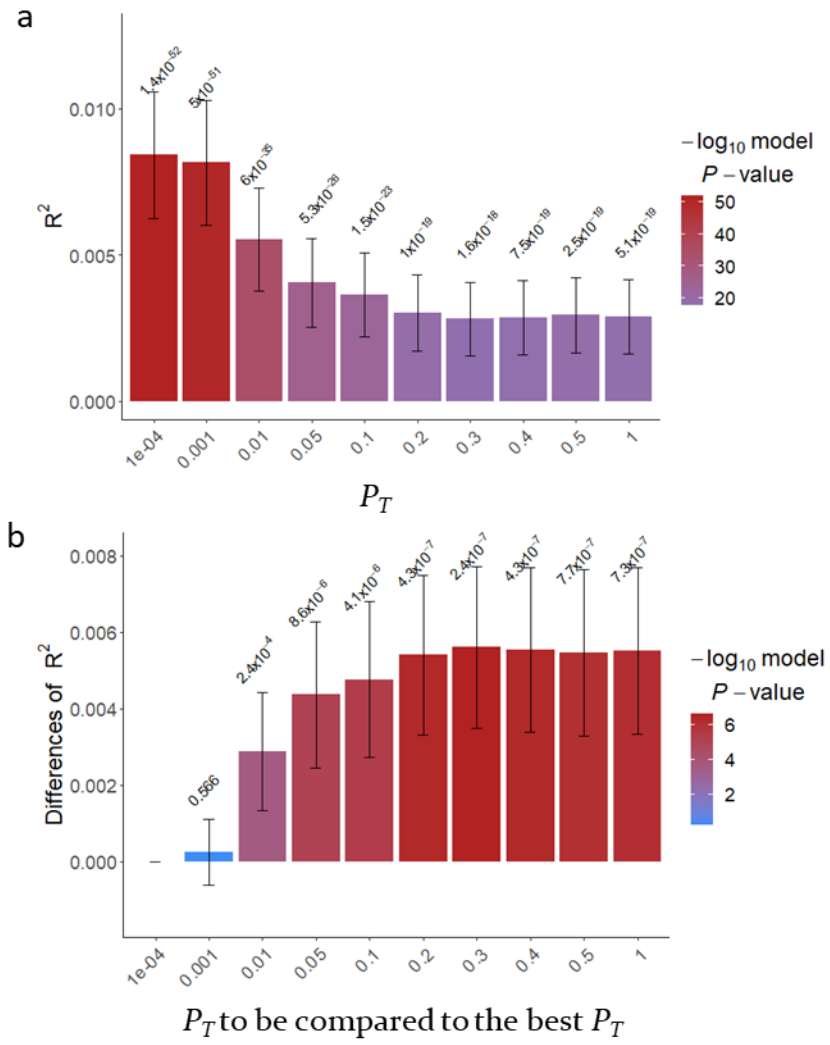

Figure 4: The predictive ability (R^2) of PGS estimated based on SNPs below the p_T when predicting cholesterol in 28,880 European samples using BBJ discovery samples (GWAS summary statistics).

a) The main bars represent R^2 values and error bars correspond 95% confidence intervals. The values above 95% CIs are p-values indicating that R^2 values are not different from zero.

b) The main bars represent the difference of R^2 values between the corresponding p_T and the best-performed p_T and error bars indicate 95% confidence intervals. The values above 95% CIs are p-values indicating the significances of differences between the pairs of R^2 values.

PGS-based genomic partitioning analyses

Genomic partitioning analyses have been widely applied [27; 29-31]. Such analysis could be useful in the PGS context. Using eq. (17) and the information matrix, $\mathcal{H}$, we can estimate the variance of the difference between, $\frac{\widehat{\beta}^2_{regu}}{p_{exp}}$ and $\frac{\widehat{\beta}^2_{non-regu}}{(1-p_{exp})}$, where $\hat{\beta}_{regu}$ and $\hat{\beta}_{non-regu}$ are the estimated regression coefficients from a multiple regression (eq. 2), and assess if the differences is significant (i.e. the coverage of the SNPs belonged to the category). For example, we partitioned the genome-wide SNPs into the regulatory (158,653) and non-regulatory regions

(3,954,947), following Gusev et al. [27], resulting 4% of SNP coverage for the regulatory region as the expectation. We simultaneously fit two sets of PGS from regulatory and non-regulatory to get $\hat{\beta}^2_{regu}$ and $\hat{\beta}^2_{non-regu}$, using a multiple regression, and assess if the difference, $\frac{\hat{\beta}^2_{regu}}{p_{exp}} - \frac{\hat{\beta}^2_{non-regu}}{(1-p_{exp})}$, is significant (eq. 18). **Figure 5** shows that the predictive ability of regulatory SNPs was significantly higher than nonregulatory SNPs (p-value = 2.6e-19 for UKBB and 8.7e-08 for BBJ) for cholesterol. In contrast, the predictive ability of regulatory SNPs was not different from the expectation (p-value = 1.0e-01 for UKBB and 8.2e-01 for BBJ) for BMI.

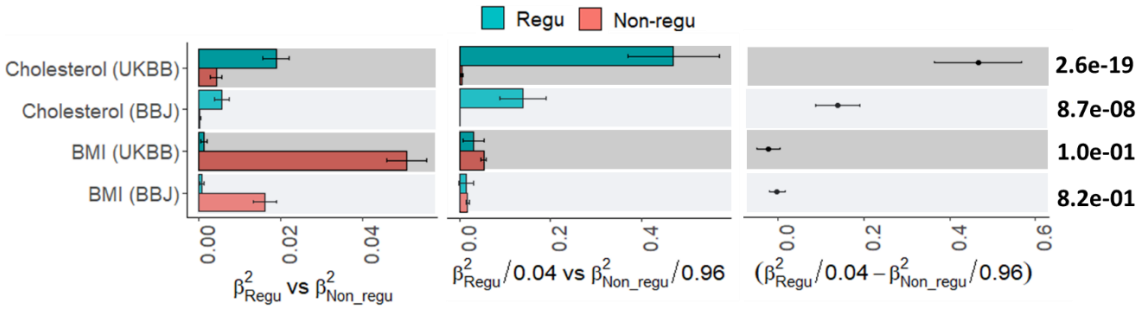

Figure 5: PGS-based genomic partitioning method to assess if the predictive ability is enriched in the regulatory region for cholesterol and BMI. Here 0.04 is the expectation for the regulatory SNPs based on the proportion of SNPs allocated to this annotation.
Left panel: The main bars represent squared regression coefficients attributable to SNPs in the regulatory region ($\hat{\beta}^2_{regu}$) and non-regulatory region ($\hat{\beta}^2_{non-regu}$), and error bars correspond 95% confidence intervals when predicting 28,880 European samples using UKBB or BBJ GWAS summary statistics.
Middle panel: The main bars represent the ratio of $\frac{\hat{\beta}^2_{regu}}{p_{exp}}$ and $\frac{\hat{\beta}^2_{non-regu}}{(1-p_{exp})}$ and error bars correspond 95% confidence intervals when predicting 28,880 European samples using UKBB or BBJ GWAS summary statistics.
Right panel: Dot points represent the difference, $(\frac{\hat{\beta}^2_{regu}}{p_{exp}} - \frac{\hat{\beta}^2_{non-regu}}{(1-p_{exp})})$ between regulatory region ($\hat{\beta}^2_{regu}$) and non-regulatory region ($\hat{\beta}^2_{non-regu}$) and error bars indicate 95% confidence intervals of the ratio differences.

Application to binary responses and ascertained case-control data

The proposed method is also explicitly verified using simulation for binary or case-control data, varying sample size and values of r^2_{y,x_1}, $r^2_{y,x_1} - r^2_{y,x_2}$, $r^2_{y,(x_1,x_2)} - r^2_{y,x_2}$ and $\frac{\hat{\beta}^2_1}{p_{exp}} - \frac{\hat{\beta}^2_2}{(1-p_{exp})}$ (**Supplementary Figures 13 – 20).** The empirical variances obtained from 10,000 simulated replicates are almost identical with the theoretical variances for the values of r^2_{y,x_1}, $r^2_{y,x_1} - r^2_{y,x_2}$,

$r^2_{y,(x_1,x_2)} - r^2_{y,x_2}$ and $\frac{\hat{\beta}_1^2}{p_{exp}} - \frac{\hat{\beta}_2^2}{(1-p_{exp})}$ when varying the sample size (**Supplementary Figures 13 – 16**) and when varying R^2 values (**Supplementary Figures 17 – 20**). In the case of ascertained case-control, a similar pattern is shown, i.e. the empirically observed variances obtained from 10,000 simulated replicates are agreed well with the theoretical variances for the values (**Supplementary Figures 21 – 24)**. This finding shows that the proposed method can be applied to test the significance of difference between predictive abilities of PGS for binary traits and ascertained case-control traits when R^2 is not very high (< 0.1). Note that the empirical and theoretical variances become disagreed when R^2 values on the observed scale are more than 0.1 for binary responses and ascertained case control (**Supplementary Figures 25 and 26**). Although R^2 value > 0.1 is not frequently observed in the current PGS studies (**Supplementary Table 2**), a careful interpretation is required for the variance of such high R^2, and we would not recommend using the theoretical approximation.

Discussion

R^2 has been widely used to measure the predictive ability of PGS [13]. However, the confidence interval of R^2 has rarely been reported, and the test statistic for the difference of two R^2 values has not been well documented. Here, we show how to get the variance of each estimated R^2 value, and covariance between two R^2 estimates (from two sets of PGS) that can be used to assess if they are significantly different to each other.

Martin et al. [18] reported that the PGS prediction accuracy is higher when discovery and target samples are from the same ancestry background, compared to when the samples are from different ancestries. However, they did not formally assess the statistical significance of the increase (no p-value provided). More importantly, they did not consider the correlation structure between predictors when they compared two PGS (in their Figure 4). We applied the proposed approach and found that the predictive ability of PGS based on UKBB discovery GWAS is significantly higher than that of PGS based on BBJ discovery GWAS, by formally deriving the 95% CI and p-value of the R^2 difference.

Many studies evaluating PGS use the p_T method [12], and report the p_T that maximises performance. This provides useful information when inferring the genetic architecture of the trait of interest and when fine-tuning p_T as a hyper-parameter in PGS methods [28; 32-34]. For such cases, it may be crucial to determine if the best-performed p_T is genuinely better than other

(adjacent) p_T or it occurs just by random chance (i.e. sampling error). Our proposed approach can formally assess statistical difference among p_T, providing 95% CI of the difference with a significance p-value.

We also derived an information matrix of squared regression coefficients in a multiple regression model, establishing a PGS-based genomic partitioning method that could test if the ratio of two squared regression coefficients is significantly deviated from its expectation given the proportion of SNPs allocated to each partition. This is analogous to the existing genomic partitioning approaches, using GREML or LDSC [27; 29-31] that may have an overfitting issue because SNP effects and genomic partitioning are estimated in the same samples.

In conclusion, we show how to estimate the variance and covariance of R^2 estimates to quantify the 95% CI and p-value of the difference and ratio when considering a pair of PGS, which is available in R package 'r2redux' (see Appendix B). We suggest that the proposed approach should be used to test the statistical significance of difference and ratio between pairs of PGS, which may help to draw a correct conclusion about the predictive ability of PGS.

Code availability

The genotype and phenotype data of the UK Biobank can be accessed through procedures described on its webpage (https://www.ukbiobank.ac.uk/) and summary statistics of BMI and cholesterol from Japanese Biobank (BBJ) can be obtained from its website (http://jenger.riken.jp/en/result)

PLINK2 version can be downloaded from https://www.cog-genomics.org/plink/

r2redux can be downloaded from (https://github.com/mommy003/r2redux_version4) and to be added in the CRAN soon (also see Appendix B).

Acknowledgements

This research is supported by the Australian Research Council (DP190100766). The R package development is supported by Cooperative Research Program for Agriculture Science and Technology Development (PJ0160992021) from the Rural Development Administration, Republic of Korea. We thank the staffs and samples of the UK Biobank and Biobank Japan for their important contributions. Our reference number approved by UK Biobank is 14575. The UK Biobank is funded by the UK Department of Health, the Medical Research Council, the

Scottish Executive, and the Wellcome Trust medical research charity. The analyses were performed using computational resources provided by the Australian Government through Gadi under the National Computational Merit Allocation Scheme (NCMAS), and HPCs (Tango and Statgen server) managed by UniSA IT. NRW acknowledges funding from the NHMRC (1113400, 1173790).

Declaration of interest

The authors declare that they have no competing interests.

Author contribution

S.H.L. and N.R.W. conceived the idea. S.H.L. derived theory and supervised the study. M.M.M performed the analysis. M.M.M and S.H.L verified the theory and analytical methods, and made the R package, with support from S.L. S.H.L and M.M.M wrote the first draft of the manuscript. N.R.W. and S.L. provided critical feedback and suggestions. All the authors discussed the results and contributed to the final manuscript.

Appendix A. The elements of Ω in eq. (7)

Following Olkin and Finn [15], each element of **Ω** in eq. (7) can be expressed as

$$var(r_{y,x_1}) = (1-\rho_{y,x_1}^2)^2/N$$

$$var(r_{y,x_2}) = (1-\rho_{y,x_2}^2)^2/N$$

$$var(r_{x_1,x_2}) = (1-\rho_{x_1,x_2}^2)^2/N$$

$$cov(r_{y,x_1}, r_{y,x_2}) = [1/2(2\rho_{x_1,x_2} - \rho_{y,x_1}\rho_{y,x_2})(1-\rho_{x_1,x_2}^2 - \rho_{y,x_1}^2 - \rho_{y,x_2}^2) + \rho_{x_1,x_2}^3]/N$$

$$cov(r_{y,x_1}, r_{x_1,x_2}) = [1/2(2\rho_{y,x_2} - \rho_{y,x_1}\rho_{x_1,x_2})(1-\rho_{x_1,x_2}^2 - \rho_{y,x_1}^2 - \rho_{y,x_2}^2) + \rho_{y,x_2}^3]/N$$

$$cov(r_{y,x_2}, r_{x_1,x_2}) = [1/2(2\rho_{y,x_1} - \rho_{y,x_2}\rho_{x_1,x_2})(1-\rho_{x_1,x_2}^2 - \rho_{y,x_1}^2 - \rho_{y,x_2}^2) + \rho_{y,x_1}^3]/N$$

Appendix B. r2redux manual

The 'r2redux' package can be used to derive test statistics for R^2 values from polygenic risk score (PRS) models (variance and covariance of R^2 values, p-value and 95% confidence intervals (CI)). For example, it can test if two sets of R^2 values from two different PRS models are significantly different to each other whether the two sets of PRS are independent or dependent. Because R^2 value is often regarded as the predictive ability of PRS, r2redux package can be useful to assess the performances of PRS methods or multiple sets of PRS based on different information sources. Furthermore, the package can derive the information matrix of $\hat{\beta}_1^2$and $\hat{\beta}_2^2$ from a multiple regression (see olkin_beta1_2 or olkin_beta_info function in the manual), which is a basis of a novel PRS-based genomic partitioning method (see r2_enrich or r2_enrich_beta function in the manual). It is recommended that the target sample size in the PGS study should be more than 2,000 for quantitative traits (Supplementary Figure 27) and more than 5,000 for binary responses or case-control studies (Supplementary Figures 28 and 29). The p-value generated from r2redux is a two-tail test. Depending on hypothesis, one-tail p-value can be obtained as the two-tail p-value divided by 2.

Installation

To use r2redux:

- `install.packages("devtools")`
- `library(devtools)`

- `devtools::install_github("mommy003/r2redux_version4")` or
- `install.packages("r2redux")` [to be added in the CRAN soon]
- `library(r2redux)`

Quick start

We illustrate the usage of r2redux using multiple sets of PRS estimated based on GWAS summary statistics from UK Biobank or Biobank Japan (reference datasets). In a target dataset, the phenotypes of target samples (y) can be predicted with PRS (a PRS model, e.g. $y = PRS + e$, where y and PRS are column-standardised [15]. Note that the target individuals should be independent from reference individuals. We can test the significant differences of the predictive ability (R^2) between a pair of PRS (see r2_diff function and example in the manual).

Data preparation

a. Statistical testing of significant difference between R^2 values for p-value thresholds: r2redux requires only phenotype and estimated PRS (from PLINK or any other software). Note that any missing value in the phenotypes and PRS tested in the model should be removed. If we want to test the significant difference of R^2 values for p-value thresholds, r2_diff function can be used with an input file that includes the following fields (also see test_ukbb_thresholds_scaled in the example directory form github (https://github.com/mommy003/r2redux_version4) and r2_diff function in the manual).

- Phenotype (y)
- PRS for p value 1 (x_1)
- PRS for p value 0.5 (x_2)
- PRS for p value 0.4 (x_3)
- PRS for p value 0.3 (x_4)
- PRS for p value 0.2 (x_5)
- PRS for p value 0.1 (x_6)
- PRS for p value 0.05 (x_7)
- PRS for p value 0.01 (x_8)
- PRS for p value 0.001 (x_9)
- PRS for p value 0.0001 (x_{10})

To get the test statistics for the difference between R^2(y~x[,v1]) and R^2(y~x[,v2]). (here we define R_1^2= R^2(y~x[,v1])) and R_2^2=R^2(y~x[,v2]))

dat=read.table("test_ukbb_thresholds_scaled") (see example files)
nv=length(dat$V1)
v1=c(1)
v2=c(2)
output=r2_diff(dat,v1,v2,nv)
r2redux output
output$var1 (variance of R_1^2)
0.0001437583
output$var2 (variance of R_2^2)
0.0001452828
output$var_diff (variance of difference between $R_1^2$and R_2^2)
5.678517e-07
output$r2_based_p (p-value for significant difference between R_1^2 and R_2^2)
0.5514562
output$mean_diff (differences between R_1^2 and R_2^2)
-0.0004488044
output$upper_diff (upper limit of 95% CI for the difference)
0.001028172
output$lower_diff (lower limit of 95% CI for the difference)
-0.001925781

b. PRS-based genomic enrichment analysis: If we want to perform some enrichment analysis (e.g., regulatory vs non_regulatory) in the PRS context to test significantly different from the expectation (4% = # SNPs in the regulatory / total # SNPs). We simultaneously fit two sets of PRS from regulatory and non-regulatory to get $\hat{\beta}^2_{regu}$ and $\hat{\beta}^2_{non-regu}$, using a multiple regression, and assess if the ratio, $\hat{\beta}^2_{regu}$/($\hat{\beta}^2_{regu}$ + $\hat{\beta}^2_{non-regu}$) and/or $\frac{\hat{\beta}^2_{regu}}{p_{exp}} - \frac{\hat{\beta}^2_{non-regu}}{(1-p_{exp})}$, are significantly different from the expectation. To test this, we need to prepare input file for r2redux that includes the following fields (e.g. test_ukbb_enrichment_choles in example directory and r2_enrich_beta function in the manual).

- Phenotype (y)
- PRS for regulatory region (x_1)
- PRS for non-regulatory region (x_2)

To get the test statistic for the ratio which is significantly different from the expectation. var($\frac{t_1}{p_{exp}} - \frac{t_2}{1-p_{exp}}$), where t_1 = $\hat{\beta}_1^2$ and t_2 = $\hat{\beta}_2^2$. β_1 and β_2 are regression coefficients from a multiple regression model, i.e. $y = x_1.\beta_1 + x_2.\beta_2 + e$, where y, x_1 and x_2 are column standardised.

dat=read.table("test_ukbb_enrichment_choles") (see example file)
nv=length(dat$V1)
v1=c(1)
v2=c(2)
expected_ratio=0.04
output=r2_enrich_beta(dat,v1,v2,nv,expected_ratio)
output
r2redux output
output$beta1_sq ($t_1$)
0.01118301
output$beta2_sq ($t_2$)
0.004980285
output$var1 (variance of t_1)
7.072931e-05
output$var2 (variance of t_2)
3.161929e-05
output$var1_2 (variance of difference between t_1 and t_2)
0.000162113
output$cov (covariance between t_1 and t_2)
-2.988221e-05
output$enrich_p2 (p-value for testing the difference between $\frac{t_1}{p_{exp}}$ and $\frac{t_2}{1-p_{exp}}$)
0.1997805
output$mean_diff (difference between $\frac{t_1}{p_{exp}}$ and $\frac{t_2}{1-p_{exp}}$)
0.2743874
output$var_diff (variance of difference, $\frac{t_1}{p_{exp}} - \frac{t_2}{1-p_{exp}}$)
0.04579649
output$upper_diff (upper limit of 95% CI for the mean difference)
0.6938296
output$lower_diff (lower limit of 95% CI for the mean difference)
-0.1450549

The r2redux manual and their example files can be downloaded from https://github.com/mommy003/r2redux_version4

References

1. Plomin, R., Haworth, C., and Davis, O.S. (2009). Common disorders are quantitative traits. Nature reviews genetics 10, 872-878.
2. Schork, N.J. (1997). Genetics of complex disease: approaches, problems, and solutions. American journal of respiratory and critical care medicine 156, S103-S109.
3. Gibson, G. (2009). Decanalization and the origin of complex disease. Nature Reviews Genetics 10, 134-140.
4. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., and Ellinor, P.T. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics 50, 1219-1224.
5. Ding, Y., Hou, K., Burch, K.S., Lapinska, S., Privé, F., Vilhjálmsson, B., Sankararaman, S., and Pasaniuc, B. (2022). Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification. Nature Genetics 54, 30-39.
6. Bilkey, G.A., Burns, B.L., Coles, E.P., Bowman, F.L., Beilby, J.P., Pachter, N.S., Baynam, G., JS Dawkins, H., Nowak, K.J., and Weeramanthri, T.S. (2019). Genomic testing for human health and disease across the life cycle: applications and ethical, legal, and social challenges. Frontiers in Public Health 7, 40.
7. Allyse, M.A., Robinson, D.H., Ferber, M.J., and Sharp, R.R. (2018). Direct-to-consumer testing 2.0: emerging models of direct-to-consumer genetic testing. In Mayo clinic proceedings. (Elsevier), pp 113-120.
8. Frerichs, F., Dingemans, K.P., and Brinkman, K. (2002). Cardiomyopathy with mitochondrial damage associated with nucleoside reverse-transcriptase inhibitors. The New England Journal of Medicine 347, 1895-1896.
9. Wray, N.R., Goddard, M.E., and Visscher, P.M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. Genome research 17, 1520-1528.
10. Wray, N.R., Yang, J., Goddard, M.E., and Visscher, P.M. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 6, e1000864.
11. Wand, H., Lambert, S.A., Tamburro, C., Iacocca, M.A., O'Sullivan, J.W., Sillari, C., Kullo, I.J., Rowley, R., Dron, J.S., and Brockman, D. (2021). Improving reporting standards for polygenic scores in risk prediction studies. Nature 591, 211-219.
12. Choi, S.W., Mak, T.S.H., and O'Reilly, P.F. (2020). A guide to performing Polygenic Risk Score analyses. Nature protocols 15, 2759.
13. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. Genomic Medicine 12, 1-11.
14. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P., and Consortium, I.S. (2009). Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder. 460, 748.
15. Olkin, I., and Finn, J.D. (1995). Correlations redux. Psychological Bulletin 118, 155.
16. Lee, S.H., Goddard, M.E., Wray, N.R., and Visscher, P.M. (2012). A better coefficient of determination for genetic profile analysis. Genetic epidemiology 36, 214-224.
17. So, H.-C., and Sham, P.C. (2017). Exploring the predictive power of polygenic scores derived from genome-wide association studies: a study of 10 complex traits. Bioinformatics 33, 886-892.
18. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nature Genetics 51, 584-591.

19. Olkin, I., and Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. Essays in probability and statistics, 235-251.
20. Olkin, I., and Finn, J.D. (1990). Testing correlated correlations. Psychological Bulletin 108, 330.
21. Wishart, J., Kondo, T., and Elderton, E. (1931). The mean and second moment coefficient of the multiple correlation coefficient, in samples from a normal population. Biometrika 22, 353-376.
22. 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.
23. Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N.E. (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. American Journal of Epidemiology 186, 1026-1034.
24. Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., and Matsuda, K. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nature genetics 49, 1458-1467.
25. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., and Matsuda, K. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nature genetics 50, 390-400.
26. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., and Daly, M.J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81, 559-575.
27. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., and Stahl, E. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. The American Journal of Human Genetics 95, 535-552.
28. Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: polygenic risk score software. Bioinformatics 31, 1466-1468.
29. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., De Andrade, M., Feenstra, B., Feingold, E., and Hayes, M.G. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nature genetics 43, 519-525.
30. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., and Wray, N.R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nature Genetics 44, 247-250.
31. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., and Farh, K. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics 47, 1228.
32. Choi, S.W., and O'Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience 8, giz082.
33. Zhao, Z., Yi, Y., Song, J., Wu, Y., Zhong, X., Lin, Y., Hohman, T.J., Fletcher, J., and Lu, Q. (2021). PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics. Genome biology 22, 1-19.

776 34. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S.,
777 Genovese, G., Loh, P.-R., Bhatia, G., and Do, R. (2015). Modeling linkage
778 disequilibrium increases accuracy of polygenic risk scores. The American Journal of
779 Human Genetics 97, 576-592.
780