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

Authors

Md. Moksedul Momin, Soohyun Lee, Naomi R. Wray, S. Hong Lee

Correspondence

cvasu.momin@gmail.com (M.M.M.), hong.lee@unisa.edu.au (S.H.L.)

R^2 is a well-established measure for the reliability of polygenic score models although its significance test is rarely considered in this context. We release an R package r2redux that allows formal statistical comparison of two polygenic score models, providing the 95% confidence interval and significance of R^2 difference.

Momin et al., 2023, The American Journal of Human Genetics *110*, 349–358
February 2, 2023

https://doi.org/10.1016/j.ajhg.2023.01.004

CellPress

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,*]

Summary

The coefficient of determination (R^2) is a well-established measure to indicate the predictive ability of polygenic scores (PGSs). 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 PGSs based on different discovery samples, the sampling covariance of R^2 is required 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 calculating PGSs in 28,880 European participants derived from UK Biobank (UKBB) and Biobank Japan (BBJ) GWAS summary statistics for cholesterol and BMI. We quantify the significantly higher predictive ability of UKBB PGSs compared to BBJ PGSs (p value 7.6e−31 for cholesterol and 1.4e−50 for BMI). A joint model of UKBB and BBJ PGSs 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). We also show that the predictive ability of regulatory SNPs is significantly enriched over non-regulatory SNPs for cholesterol (p value 8.9e−26 for UKBB and 3.8e−17 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 PGSs, which may help to draw a correct conclusion about the comparative predictive ability of PGSs.

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 (GWASs) have provided an opportunity to estimate genetic profile or polygenic scores (PGSs) that represent 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., PGSs) 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) is a useful measure to quantify the reliability of the PGS. Note that R^2 is equivalent to the squared regression coefficient if the dependent and explanatory variables are column standardized.[15]

Previously, we introduced a 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 estimates depend on the proportion of affected individuals in the sample. In contrast, R^2 on the liability scale does not depend on the proportion of cases in the sample but does require an estimate of the lifetime population prevalence of the disease.

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 PGSs, using R^2.[12,13,17,18] However, the variance of R^2[15] has been rarely studied especially in the context of PGSs although it is the crucial parameter for estimation of confidence intervals (CI) of R^2. Furthermore, estimates of the covariance between a pair of R^2 values (e.g., from two sets of PGSs) are necessary to assess whether they are significantly different from each other, or if the ratio of two R^2 values significantly deviates from the expectation. This significance test for the difference or ratio is important when comparing two or multiple sets of PGSs that are derived from different sets of SNPs, e.g., genomic partitioning, genome-wide association p value thresholds (p_T) analysis, or PGSs based on pathway subsets.[19,20]

[1]Australian Centre for Precision Health, 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), Cheonan, South Korea; [6]Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia; [7]Queensland Brain Institute, University of Queensland, Brisbane, QLD, Australia
*Correspondence: cvasu.momin@gmail.com (M.M.M.), hong.lee@unisa.edu.au (S.H.L.)
https://doi.org/10.1016/j.ajhg.2023.01.004.

In this study, we use R^2 measures and their variance-covariance matrix to assess whether the predictive abilities of PGSs based on different sources are significantly different from 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 PGSs. We also derive the variance and covariance matrix (i.e., information matrix) of squared regression coefficients in a multiple regression model, testing whether 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 compare PGSs calculated in 28,880 European individuals using UK Biobank (UKBB) and Biobank Japan (BBJ) GWAS summary statistics for cholesterol and BMI.

Material and methods

We used data from the UK Biobank (https://www.ukbiobank.ac.uk), the scientific protocol of which has been reviewed and approved by the Northwest Multi-center Research Ethics Committee, National Information Governance Board for Health & Social Care, and Community Health Index Advisory Group. UK Biobank has obtained informed consent from all participants. Our access to the UK Biobank data was under the reference number 14575.

Publicly available GWAS summary statistics of Biobank Japan (BBJ)[21,22] were used, following BBJ's guidelines (http://jenger.riken.jp/en/result). The research ethics approval of this study has been obtained from the University of South Australia Human Research Ethics Committee.

PGS models

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

A PGS model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad \text{(Equation 1)}$$

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

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

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

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

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

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

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

$$r^2_{y,(x_1,x_2)} = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N} y_i^2} = \frac{\sum_{i=1}^{N}\widehat{y}_i^2}{N} = \widehat{\beta}_1^2 + \widehat{\beta}_2^2 + 2r_{x_1,x_2}\widehat{\beta}_1\widehat{\beta}_2. \qquad \text{(Equation 4)}$$

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

$$r^2_{y,x_1} = \frac{\sum_{i=1}^{N}\widehat{y}_i^2}{N} = \widehat{\beta}_1^2 \text{ with } \mathbf{X} = [\mathbf{x_1}]$$

or

$$r^2_{y,x_2} = \frac{\sum_{i=1}^{N}\widehat{y}_i^2}{N} = \widehat{\beta}_2^2 \text{ with } \mathbf{X} = [\mathbf{x_2}].$$

It is noted that $r^2_{y,(x_1,x_2)}$, r^2_{y,x_1}, or r^2_{y,x_2} is an estimate of parameter $\rho^2_{y,(x_1,x_2)}, \rho^2_{y,x_1}$, *or* ρ^2_{y,x_2}, and each estimate has a sampling variance.

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 × $(M+2\lambda)$ where $\lambda = \frac{N \times R^2}{(1-R^2)^2}$ is the non-centrality parameter. For example, the variance of the transformed value for r^2_{y,x_1} is

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

Therefore,

$$var\left(r^2_{y,x_1}\right) = var(\widehat{\beta}_1^2) = 2var(\widehat{\beta}_1)^2(M+2\lambda) \qquad \text{(Equation 5)}$$

where $var(\widehat{\beta}_1) = 1/N \cdot (1 - \rho^2_{y,x_1})^2$, M = 1, and ρ^2_{y,x_1} is the squared correlation in the population and can be approximated as $\rho^2_{y,x_1} \approx r^2_{y,x_1}$.[23,24]

In a similar manner, Equation 5 can be extended to multiple explanatory variables as

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

that is, Equation 6 is a generalized form of Equation 5.

Wishart[25] introduced a formula to obtain the variance of R^2 (also see Stuart and Ord[26] and Olkin and Finn[15]) as

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

which provides an equivalent estimate as in Equation 6. Wishart[25] derived his formula of the variance of R^2 based on the hypergeometric series that has been used in the literature including Olkin

and Finn.[15] We introduce Equation 6 derived based on the transformation of a non-central χ^2 distribution. Both Equation 6 and Wishart equation provide identical estimates of the variance of R^2 (Figure S1). 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., $f(r_{y,x_1}, r_{y,x_2}, r_{x_1,x_2})$, the delta method approximates the variance of the difference as

$$var(f) = \boldsymbol{\theta}'\boldsymbol{\Omega}\boldsymbol{\theta} \qquad \text{(Equation 7)}$$

where

$$\boldsymbol{\theta}' = \left(\frac{\partial f}{\partial r_{y,x_1}}, \frac{\partial f}{\partial r_{y,x_2}}, \frac{\partial f}{\partial r_{x_1,x_2}}\right) \qquad \text{(Equation 8)}$$

is the derivatives of f with respect to the correlations 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 Supplemental Note A).

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

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

The variance of R^2 difference can be written as

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

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 Equation 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

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

and the 95% confidence interval is

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

When comparisons are made between R^2 values based on two sets of PGSs ($\mathbf{x_1}$ and $\mathbf{x_2}$), the sampling covariance of R^2 is required, which is explicitly used in Equations 7 and 9. If the sampling covariance ignored, the test statistics can be biased (Figures S2 and S3).

R^2 *difference when using nested models*

When using nested models, the variance of R^2 difference can be written as

$$var\left(r^2_{y,(x_1,x_2)} - r^2_{y,x_2}\right) \text{ with } f(r_{y,x_1}, r_{y,x_2}, r_{x_1,x_2}) = r^2_{y,(x_1,x_2)} - r^2_{y,x_2} = \widehat{\beta}^2_1 + \widehat{\beta}^2_2 + 2r_{x_1,x_2}\widehat{\beta}_1\widehat{\beta}_2 - r^2_{y,x_2} \qquad \text{(Equation 11)}$$

where $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are the estimated regression coefficients from a multiple regression (Equation 2), calculated from $\boldsymbol{\Sigma}$ (see Equations 2, 3, and 4). Again, the derivative with respect to each of the correlations can be obtained for this function (Equation 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 Equation 11 from random samples in the population follows a non-central chi-squared distribution with a non-centrality parameter $= N \times \frac{r^2_{y,(x_1,x_2)} - r^2_{y,x_2}}{\left(1 - r^2_{y,(x_1,x_2)} - r^2_{y,x_2}\right)^2}$. 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

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

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] and the 95% confidence interval is

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

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

R^2 *difference when using two independent sets of PGSs*

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

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

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

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

The p value for the significance test of the difference can be derived from

$$\frac{\left(r^2_{y_1,x_1} - r^2_{y_2,x_2}\right)^2}{var\left(r^2_{y_1,x_1} - r^2_{y_2,x_2}\right)} \sim \chi^2_1$$

and the 95% confidence interval[15] is

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

(Equation 15)

PGS-based genomic partitioning analysis

It is of interest to test whether a set of PGSs based on a genomic region of interest (or a pathway-based SNP subset) can better predict the phenotypes, compared to the rest of genomic regions. The proportion of the coefficient of determination explained by x_1 can be estimated as $\widehat{\beta}^2_1/r^2_{y,(x_1,x_2)}$ from a multiple regression, $y = x_1 + x_2 + e$, where x_1 is the PGS of a genomic region of interest and x_2 is the PGS of the rest of genomic regions. The expected proportion of the coefficient of determination explained by x_1 can be calculated from prior information, referred to as p_{exp} = # SNPs used for PGS1/total # SNPs. We are interested in testing whether the value of $\widehat{\beta}^2_1/r^2_{y,(x_1,x_2)}$ is significantly different from its expectation, p_{exp}, which requires to estimate the sampling variance of $\widehat{\beta}^2_1/r^2_{y,(x_1,x_2)}$, using Equation 7. The variance of the proportion can be written as

$$var\left(\widehat{\beta}^2_1 \Big/ r^2_{y,(x_1,x_2)}\right) \text{ with } f\left(r_{y,x_1}, r_{y,x_2}, r_{x_1,x_2}\right) = \widehat{\beta}^2_1 \Big/ r^2_{y,(x_1,x_2)}$$

(Equation 16)

where $\widehat{\beta}_1$ is the estimated regression coefficient of x_1, calculated from $\boldsymbol{\Sigma}$ (Equation 3), and $r^2_{y,(x_1,x_2)} = \widehat{\beta}^2_1 + \widehat{\beta}^2_2 + 2r_{x_1,x_2}\widehat{\beta}_1\widehat{\beta}_2$ is the coefficient of determination. 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 Equation 8. This variance can be used to obtain the significance and 95 CI of the observed proportion of the coefficient of determination.

Analogous to Equation 9, the values of $\frac{\widehat{\beta}_1}{r^2_{y,(x_1,x_2)}} - p_{exp}$ 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

$$\frac{\left[\left(\frac{\widehat{\beta}^2_1}{r^2_{y,(x_1,x_2)}} - p_{exp}\right)\right]^2}{var\left(\frac{\widehat{\beta}^2_1}{r^2_{y,(x_1,x_2)}} - p_{exp}\right)} \sim \chi^2_1$$

The 95% confidence interval of the ratio is

$$\left[\left(\frac{\widehat{\beta}^2_1}{r^2_{y,(x_1,x_2)}} - p_{exp}\right) - 1.96\sqrt{var\left(\frac{\widehat{\beta}^2_1}{r^2_{y,(x_1,x_2)}} - p_{exp}\right)}, \left(\frac{\widehat{\beta}^2_1}{r^2_{y,(x_1,x_2)}} - p_{exp}\right) + 1.96\sqrt{var\left(\frac{\widehat{\beta}^2_1}{r^2_{y,(x_1,x_2)}} - p_{exp}\right)}\right]$$

(Equation 17)

In addition, the package, r2redux, can provide $var(\widehat{\beta}^2_1)$, $var(\widehat{\beta}^2_2)$, and $var(\widehat{\beta}^2_1 - \widehat{\beta}^2_2)$, i.e., the information matrix of the squared regression coefficients (see Supplemental Note B) that is useful when comparing the actual values of $\widehat{\beta}^2_1$ and $\widehat{\beta}^2_2$.

It is noted that the delta method employed in this study is a well-established approach to derive the distribution of a function of an asymptotically normal variable.[27] Following Olkin and Finn,[15] we used the delta method to derive the variances of R^2 and their difference as a function of regression coefficients (Equations 7, 8, 9, 11, and 16). We explicitly checked that the regression coefficients are asymptotically normal, using a realistic correlation structure among variables (Figures S4–S6).

Data

The UK Biobank is a large-scale biomedical database that comprises 0.5 million individuals who had been recruited between 2006 and 2010; their age ranged between 40 and 69 years.[28,29] The data consist 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 sex between reported and inferred from genotypic information, genotype call rate < 0.95, or 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, or Hardy-Weinberg equilibrium p value $<10^{-7}$. We also applied a relatedness cut-off QC (>0.05) so that there was no high pairwise relatedness among individuals. After QC, 288,792 individuals and 7,701,772 SNPs were retained.

Discovery GWAS data

Ninety percent of the individuals from the 288,792 QCed individuals were randomly selected as discovery samples (n = 259,912 to generate GWAS summary statistics (UKBB hereafter) for the 7,701,772 SNPs.. 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 center, and the first 10 ancestry principal components 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[21] (n = 158,284) and cholesterol[22] (n = 128,305) for 5,961,601 SNPs.

Target data

Ten percent of the individuals from the 288,792 QCed individuals were randomly selected as an independent target dataset (n = 28,880) 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 center, and the first 10 principal components using a linear regression. The pre-adjusted phenotypes were correlated with PGSs estimated in the following step. For each trait, we used the UKBB and BBJ GWAS summary statistics to estimate two sets of PGSs (UKBB PGSs vs. BBJ PGSs for the 28,880 target individuals), using PLINK2 (https://www.cog-genomics.org/plink/2.0/) with the score function.[30] Then, we estimated the correlation between the PGS and pre-adjusted phenotypes to obtain R^2 values in the PGS analyses.

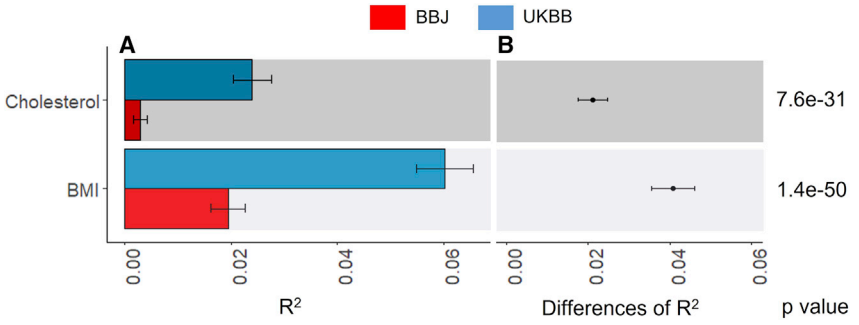

Figure 1. The predictive ability (R^2) of PGSs when predicting 28,880 European individuals using UKBB or BBJ discovery GWAS dataset

(A) 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 PGSs. (B) 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.

Functional annotation of the genome

We annotated the genome using pre-defined functional categories (regulatory vs. non-regulatory genomic regions).[31] Regulatory region includes SNPs from coding regions, untranslated regions (UTRs), 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 PGSs (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 Figures S7–S15). 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 standardized 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 $\frac{\widehat{\beta}^2_1}{r^2_{y,(x_1,x_2)}} - p_{exp}$ were obtained over 10,000 replicates, which were compared to the theoretical variances estimated using Equations 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 $\frac{\widehat{\beta}^2_1}{r^2_{y,(x_1,x_2)}} - p_{exp}$ (Figures S7–S15). 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}, $r^2_{y,x_1} - r^2_{y,x_2}$, $r^2_{y,(x_1,x_2)} - r^2_{y,x_2}$, $\frac{\widehat{\beta}^2_1}{r^2_{y,(x_1,x_2)}} - p_{exp}$ when varying the sample size (Figures S7–S10) and when varying R^2 values (Figures S11–S14). When considering two independent PGSs, the theoretical and empirical variances are also agreed well (Figure S15).

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 Equations 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 PGSs 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 PGSs are not independent. Furthermore, the two PGS models with UKBB and BBJ are not nested to each other, so the likelihood ratio test could not be used either. For this problem, we used Equations 9 and 10 to obtain the variance, 95% CI (0.0175–0.0247), and p value (7.6e−31) of the R^2 difference, accounting for the dependency between UKBB and BBJ PGSs, 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).

It is also interesting to whether BBJ PGSs provides a significant improvement in the predictive ability, in addition to UKBB PGSs, when predicting the 28,880 European target samples. Figure 2 compares R^2 value with each UKBB or BBJ PGSs to R^2 value from a joint model fitting UKBB and BBJ PGSs simultaneously. Using Equations 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 PGSs contributed to a significant improvement for cholesterol. Similarly, BBJ PGSs improved the predictive ability significantly (p value = 1.3e−28) for BMI. As expected, excluding UKBB PGSs from the joint model substantially decreased the prediction accuracy (p value = 1.6e−136 for cholesterol and 3.0e−308 for BMI).

R^2 difference when using two independent sets of PGSs: male vs. female

We were also interested in testing whether the PGSs 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 PGSs,

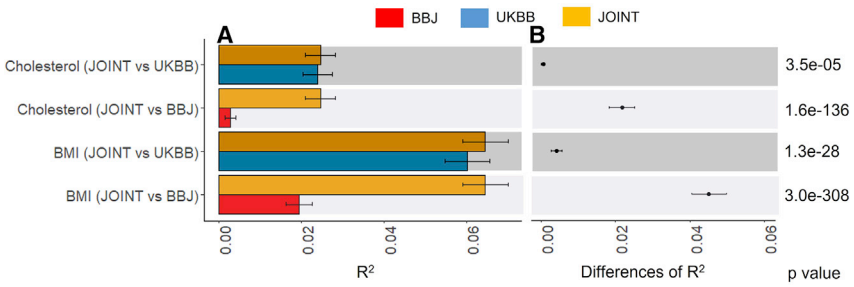

Figure 2. The predictive ability (R^2) of a PGS model based on UKBB or BBJ discovery dataset, compared to the joint model of both UKBB and BBJ when predicting 28,880 European individuals

(A) 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 PGSs, i.e., UKBB and BBJ PGSs. In addition, a joint model fitting both UKBB and BBJ PGSs was compared.

(B) 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.

so the variance of R^2 difference is simply the sum of the variances of each R^2 value, which can be obtained from Equation 5 or 6. Figure S16 shows that there was no significant difference between male and female PGSs in their predictive ability for cholesterol and BMI whether using UKBB or BBJ discovery GWAS dataset.

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

PGSs also have been widely used to determine which p_T provides the highest prediction accuracy, for example, using PGS software such as PLINK.[30,32] However, there is a lack of test statistics that can assess whether the predictive ability of the best-performing 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-performing p_T. Using Equations 9 and 10, we assessed the significance of difference between the best-performing p_T and each of the other p_T (Figure 3B). From this analysis, we found that the best-performing p_T was not significantly different from p_T ranging between 0.1 and 1, but significantly different from $p_{T<0.05}$ (Figure 3B). When using the 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 (Figure S17B).

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 the UKBB discovery GWAS dataset (Figure S18A). 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 (Figure S18B).

PGS-based genomic partitioning analyses

Genomic partitioning analyses have been widely applied.[31,33–35] Such analysis could be useful in the PGS context. Using Equation 16, we can estimate the variance of the $\frac{\widehat{\beta}^2_{regu}}{R^2}$ where $\widehat{\beta}_{regu}$ is the estimated regression coefficient from a multiple regression (Equation 2), and assess whether the observed proportion ($\frac{\widehat{\beta}^2_{regu}}{R^2}$) is significant different from p_{exp} (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 (3,954,947) regions, following Gusev et al.,[31] resulting $p_{exp} = 4\%$ of SNP coverage for the regulatory region as the expectation. We simultaneously fit two sets of PGSs from regulatory and non-regulatory regions to get $\widehat{\beta}^2_{regu}$ and $\widehat{\beta}^2_{non-regu}$, using a multiple regression, then assess whether the value of $\frac{\widehat{\beta}^2_{regu}}{R^2} - p_{exp}$ is significantly different from zero (Equation 17). Figure 5 shows that the predictive ability of regulatory SNPs was significantly higher than the expectation (p value = 8.9e−26 for UKBB and 3.8e−17 for BBJ) for cholesterol. In contrast, the predictive ability of regulatory SNPs was not better than the expectation for BMI (Figure 5).

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 $\frac{\widehat{\beta}^2_1}{R^2} - p_{exp}$ (Figures S19–S26). 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 $\frac{\widehat{\beta}^2_1}{R^2} - p_{exp}$ when varying the sample size (Figures S19–S22) and when varying R^2 values (Figures S23–S26). 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 (Figures S27–S30). This finding shows that the proposed method can be applied to test the significance of difference between predictive abilities of PGSs for binary traits and ascertained case-control traits when R^2 is not

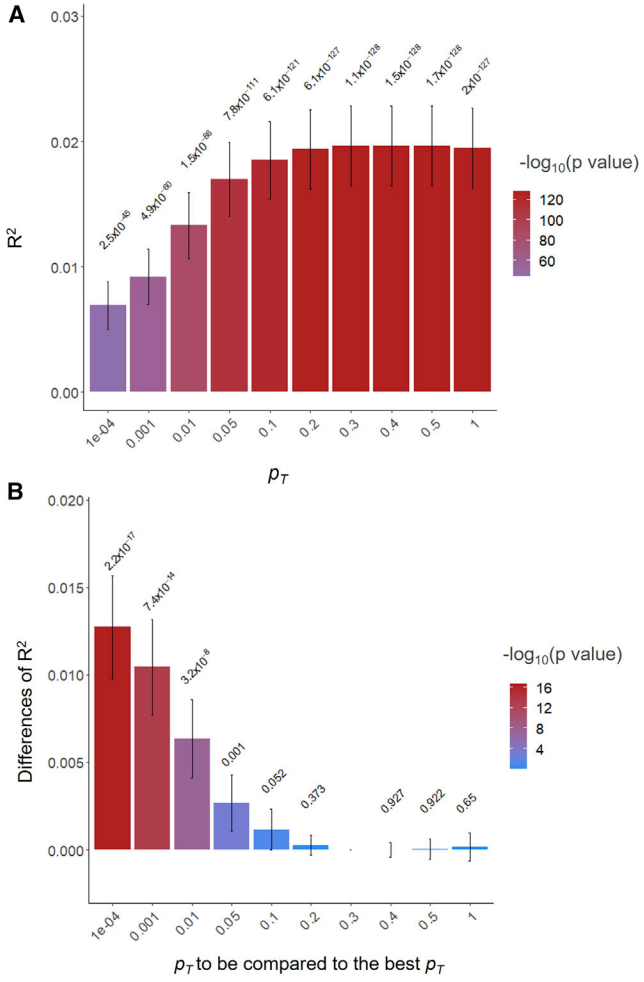

Figure 3. The predictive ability (R^2) of PGSs estimated based on SNPs below p_T when predicting BMI in 28,880 European samples using BBJ discovery samples (GWAS summary statistics)
(A) The main bars represent R^2 values and error bars correspond 95% confidence intervals. The values above 95% CIs are 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-performing p_T and error bars indicate 95% confidence intervals. The values above 95% CIs are p values indicating the significance of the difference between the pairs of R^2 values.

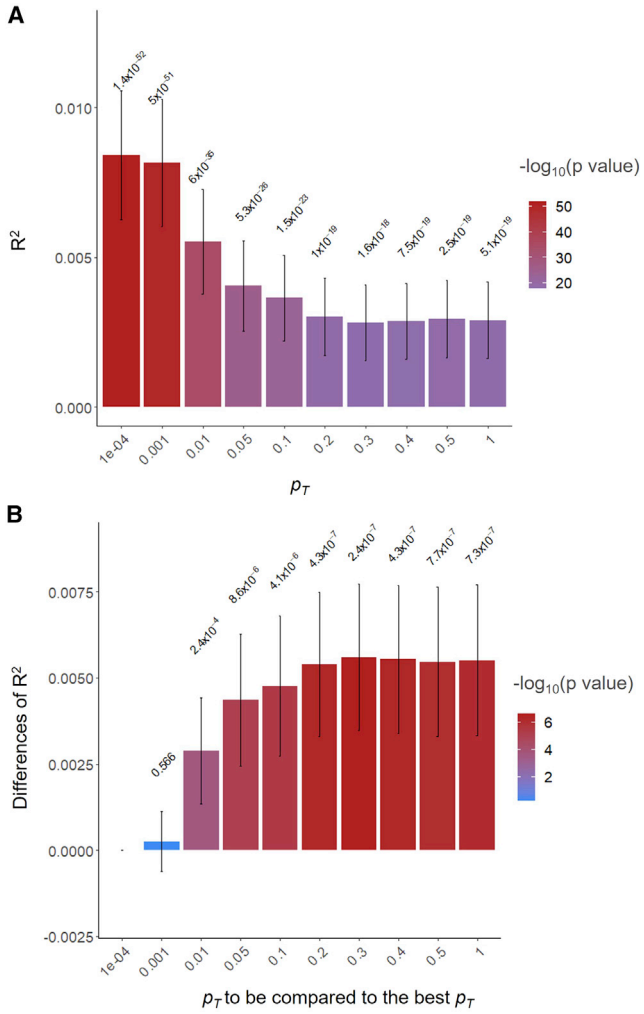

Figure 4. The predictive ability (R^2) of PGSs estimated based on SNPs below the p_T when predicting cholesterol in 28,880 European samples using BBJ discovery samples (GWAS summary statistics)
(A) The main bars represent R^2 values and error bars correspond 95% confidence intervals. The values above 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-performing p_T and error bars indicate 95% confidence intervals. The values above 95% CIs are p values indicating the significance of the difference between the pairs of R^2 values.

very high (<0.1). Note that the empirical and theoretical variances diverge when R^2 values on the observed scale are more than 0.1 for binary responses and ascertained case control (Figures S31 and S32). Although R^2 value > 0.1 is not frequently observed in the current PGS studies (Table S2), 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 PGSs.[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 PGSs) that can be used to assess whether they are significantly different from 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 PGSs. We applied the proposed approach and found that the predictive ability of PGSs

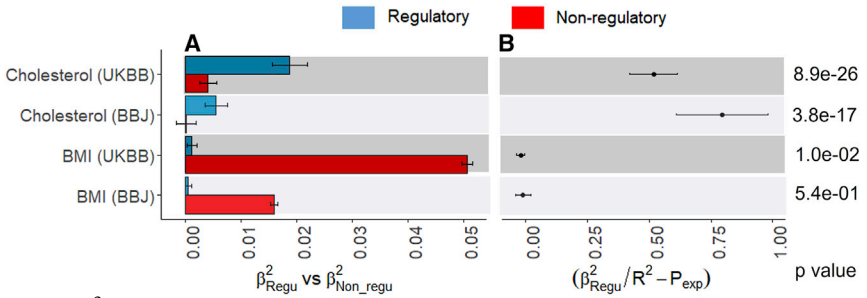

Figure 5. PGS-based genomic partitioning method to assess whether the predictive ability is enriched in the regulatory region for cholesterol and BMI

Here $p_{exp} = 0.04$ is the expectation for the regulatory SNPs based on the proportion of SNPs allocated to this annotation.

(A) The main bars represent squared regression coefficients attributable to SNPs in the regulatory region ($\widehat{\beta}^2_{regu}$) and non-regulatory region ($\widehat{\beta}^2_{non-regu}$), and error bars correspond to 95% confidence intervals when predicting 28,880 European samples using UKBB or BBJ GWAS summary statistics.

(B) Dot points represent the difference between the observed and expected proportions ($\frac{\widehat{\beta}^2_{regu}}{R^2} - p_{exp}$) and error bars indicate 95% confidence intervals of the difference.

based on UKBB discovery GWASs is significantly higher than that of PGSs based on BBJ discovery GWASs, by formally deriving the 95% CI and p value of the R^2 difference.

Many studies evaluating PGSs use the p_T method[12] and report the p_T that maximizes 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.[32,36–38] For such cases, it may be crucial to determine if the best-performing p_T is genuinely better than other (adjacent) p_T or if it occurs just by random chance (i.e., sampling error). For example, in Figure 3, the best-performing p_T is 0.3 (the set of SNPs with $p_T \leq 0.3$), which is, however, not statistically different from $p_T \leq 0.2$ or ≤ 0.1. Note that the set of SNPs with $p_T \leq 0.1$ is nested within SNPs with $p_T \leq 0.3$, meaning that the additional SNPs in the latter would not significantly improve the prediction accuracy. Therefore, $p_T \leq 0.1$ should be used instead of the $p_T \leq 0.3$ as the former is a more parsimonious model than the latter. 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 whether 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[31,33–35] 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 comparing two PGSs, which is available in R package r2redux (see Supplemental Note B). We suggest that the proposed approach should be used to test the statistical significance of difference and ratio between pairs of PGSs, which may help to draw a correct conclusion about the predictive ability of PGSs.

Data and 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/). r2redux can be downloaded from (https://github.com/mommy003/r2redux) or from CRAN [install.packages("r2redux") in R] (also see Supplemental Note B).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.01.004.

Acknowledgments

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 (PJ01609901) 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. N.R.W. acknowledges funding from the NHMRC (1113400, 1173790), Australia.

Author contributions

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.

Declaration of interests

The authors declare that they have no competing interests.

Received: June 8, 2022
Accepted: January 5, 2023
Published: January 25, 2023

References

1. Plomin, R., Haworth, C.M.A., and Davis, O.S.P. (2009). Common disorders are quantitative traits. Nat. Rev. Genet. *10*, 872–878.
2. Schork, N.J. (1997). Genetics of complex disease: approaches, problems, and solutions. Am. J. Respir. Crit. Care Med. *156*, S103–S109.
3. Gibson, G. (2009). Decanalization and the origin of complex disease. Nat. Rev. Genet. *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., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *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. Nat. Genet. *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. Front. Public Health *7*, 40.
7. Allyse, M.A., Robinson, D.H., Ferber, M.J., and Sharp, R.R. (2018). In Direct-to-consumer Testing 2.0: Emerging Models of Direct-To-Consumer Genetic Testing, *1* (Elsevier), pp. 113–120.
8. Frerichs, F.C.P., Dingemans, K.P., and Brinkman, K. (2002). Cardiomyopathy with mitochondrial damage associated with nucleoside reverse-transcriptase inhibitors. N. Engl. J. Med. *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 Res. *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., Brockman, D., et al. (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. Nat. Protoc. *15*, 2759–2772.
13. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. Genome Med. *12*. 44–11.
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. Psychol. Bull. *118*, 155–164.
16. 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.
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. Nat. Genet. *51*, 584–591.
19. Choi, S., Garcia-Gonzalez, J., Ruan, Y., Wu, H., Johnson, J., Hoggart, C., and O'Reilly, P. (2021). The power of pathway-based polygenic risk scores. Research Square. https://doi.org/10.21203/rs.3.rs-643696/v1.
20. Li, J., Chaudhary, D.P., Khan, A., Griessenauer, C., Carey, D.J., Zand, R., and Abedi, V. (2021). Polygenic risk scores augment stroke subtyping. Neurol. Genet. *7*, e560.
21. Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat. Genet. *49*, 1458–1467.
22. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.
23. Olkin, I., and Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. Essays in probability and statistics, 235–251.
24. Olkin, I., and Finn, J.D. (1990). Testing correlated correlations. Psychol. Bull. *108*, 330–333.
25. Wishart, J. (1931). The mean and second moment coefficient of the multiple correlation coefficient, in samples from a normal population. Biometrika *22*, 353–361.
26. Stuart, A., and Ord, J.K. (1991). Kendall's Advanced Theory of Statistics*Vol 2*.
27. Ver Hoef, J.M. (2012). Who invented the delta method? Am. Statistician *66*, 124–127.
28. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
29. 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. Am. J. Epidemiol. *186*, 1026–1034.
30. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
31. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.
32. Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: polygenic risk score software. Bioinformatics *31*, 1466–1468.
33. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., De Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.
34. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Schizophrenia Psychiatric Genome-Wide Association Study Consortium

PGC-SCZ; International Schizophrenia Consortium ISC; and Molecular Genetics of Schizophrenia Collaboration MGS, Sullivan, P.F., Goddard, M.E., Keller, M.C., et al. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. *44*, 247–250.

35. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
36. Choi, S.W., and O'Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. GigaScience *8*, giz082.
37. 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 Biol. *22*, 257–319.
38. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592.

The American Journal of Human Genetics, Volume *110*

Supplemental information

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

Md. Moksedul Momin, Soohyun Lee, Naomi R. Wray, and S. Hong Lee

Supplemental Note A: The elements of Ω in eq. (7)

Following Olkin and Finn [1], each element of **Ω** in eq. (7) can be expressed as

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

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

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

Supplemental Note B: r2redux manual

The 'r2redux' package can be used to derive test statistics for R^2 values from polygenic risk score (PGS) models (variance and covariance of R^2 values, p value and 95% confidence intervals (CI)) (see manual https://cran.r-project.org/web/packages/r2redux/r2redux.pdf). For example, it can test if two sets of R^2 values from two different PGS models are significantly different to each other whether the two sets of PGS are independent or dependent. Because R^2 value is often regarded as the predictive ability of PGS, r2redux package can be useful to assess the performances of PGS methods or multiple sets of PGS 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 PGS-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 (Figure S27) and more than 5,000 for binary responses or case-control studies (Figures S28 and S29). The p value generated from the r2redux package provides two types of p values (for one- and two-tailed test) unless the comparison is for nested models (e.g. $y = PGS_1 + PGS_2 + e$ vs. $y = PGS_2 + e$) where the R^2 of the full model is expected to be always higher than the reduced model. When there are multiple covariates (e.g. age, sex and other demographic variables), the phenotypes can be adjusted for the covariates, and pre-adjusted phenotypes (residuals) should be used in the r2redux.

Installation

To use r2redux:

```
install.packages("r2redux")
library(r2redux)
```

or

```
install.packages("devtools")
library(devtools)
devtools::install_github("mommy003/r2redux")
library(r2redux)
```

Quick start

We illustrate the usage of r2redux using multiple sets of PGS 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 PGS (a PGS model, e.g. $y = PGS + e$, where y and PGS are column-standardised [1]. 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 PGS (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 PGS (from PLINK or any other software). Note that any missing value in the phenotypes and PGS 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) or read dat1 file embedded within the package and r2_diff function in the manual (https://cran.r-project.org/web/packages/r2redux/r2redux.pdf).

- Phenotype (y)
- PGS for p value 1 (x_1)
- PGS for p value 0.5 (x_2)
- PGS for p value 0.4 (x_3)

- PGS for p value 0.3 (x_4)
- PGS for p value 0.2 (x_5)
- PGS for p value 0.1 (x_6)
- PGS for p value 0.05 (x_7)
- PGS for p value 0.01 (x_8)
- PGS for p value 0.001 (x_9)
- PGS 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) or
dat=dat1 (this example embedded within the package)
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)
0.0001436128
output$var2 (variance of R_2^2)
0.0001451358
output$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. PGS-based genomic enrichment analysis: If we want to perform some enrichment analysis (e.g., regulatory vs non_regulatory) in the PGS context to test significantly different from the expectation (p_{exp}= # SNPs in the regulatory / total # SNPs = 4%). We simultaneously fit two sets of PGS from regulatory and non-regulatory to get $\hat{\beta}_{regu}^2$ and $\hat{\beta}_{non-regu}^2$, using a multiple regression, and assess if the ratio, $\frac{\hat{\beta}_1^2}{r_{y,(x_1,x_2)}^2}$ are significantly different from the expectation, p_{exp}. 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 or read dat2 file embedded within the package and r2_enrich_beta function in the manual).

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

To get the test statistic for the ratio which is significantly different from the expectation. var($\hat{\beta}_1^2/r^2_{y,(x_1,x_2)}$), where $\hat{\beta}_1^2$ is the squared regression coefficient of x_1 from a multiple regression model, i.e. $y = x_1\beta_1 + x_2\beta_2 + e$, and $r^2_{y,(x_1,x_2)}$ is the coefficient of determination of the model. It is noted that y, x_1 and x_2 are column standardised (mean 0 and variance 1).

dat=read.table("test_ukbb_enrichment_choles") (see example file) or
dat=dat2 (this example data is embedded within the package)
nv=length(dat$V1)
v1=c(1)
v2=c(2)
dat=dat2
nv=length(dat$V1)
v1=c(1)
v2=c(2)
output=r2_beta_var(dat,v1,v2,nv)
r2redux output
output$beta1_sq ($\hat{\beta}_1^2$)
0.01118301
output$beta2_sq ($\hat{\beta}_2^2$)
0.004980285
output$var1 (variance of $\hat{\beta}_1^2$)
7.072931e-05
output$var2 (variance of $\hat{\beta}_2^2$)
3.161929e-05
output$var1_2 (variance of difference between $\hat{\beta}_1^2$and $\hat{\beta}_2^2$)
0.000162113
output$cov (covariance between $\hat{\beta}_1^2$and $\hat{\beta}_2^2$)
-2.988221e-05
output$upper_beta1_sq (upper limit of 95% CI for $\hat{\beta}_1^2$)
0.03037793
output$lower_beta1_sq (lower limit of 95% CI for $\hat{\beta}_1^2$)
-0.00123582
output$upper_beta2_sq (upper limit of 95% CI for $\hat{\beta}_2^2$)
0.02490076
output$lower_beta2_sq (lower limit of 95% CI for $\hat{\beta}_2^2$)
-0.005127546

dat=dat2 (this example data is embedded within the package)
nv=length(dat$V1)
v1=c(1)

v2=c(2)
expected_ratio=0.04
output=r2_enrich_beta(dat,v1,v2,nv,expected_ratio)
r2redux output
output$beta1_sq ($\hat{\beta}_1^2$)
0.01118301
output$beta2_sq ($\hat{\beta}_2^2$)
0.004980285
output$ratio1 ($\hat{\beta}_1^2/R^2$)
0.4392572
output$ratio2 ($\hat{\beta}_2^2/R^2$)
0.1956205
output$ratio_var1 (variance of ratio 1)
0.08042288
output$ratio_var2 (variance of ratio 2)
0.0431134
output$upper_ratio1 (upper limit of 95% CI for ratio 1)
0.9950922
output$lower_ratio1 (lower limit of 95% CI for ratio 1)
-0.1165778
output$upper_ratio2 upper limit of 95% CI for ratio 2)
0.6025904
output$lower_ratio2 (lower limit of 95% CI for ratio 2)
-0.2113493
output$enrich_p1 (two tailed p value for $\hat{\beta}_1^2/R^2$is significantly different from exp1)
0.1591692
output$enrich_p1_one_tail (one tailed p value for $\hat{\beta}_1^2/R^2$ is significantly different from exp1)
0.07958459
output$enrich_p2 (two tailed p value for $\hat{\beta}_2^2$/R2 is significantly different from (1-exp1))
0.000232035
output$enrich_p2_one_tail (one tailed p value for $\hat{\beta}_2^2/R^2$ is significantly different from (1-exp1))
0.0001160175

A code for an additional unit test is available in “r2redux/tests/testthat/” directory

The r2redux manual (https://cran.r-project.org/web/packages/r2redux/r2redux.pdf) and their example files can be downloaded from https://github.com/mommy003/r2redux or from CRAN [install.packages("r2redux") in R]**.**

Supplemental figures

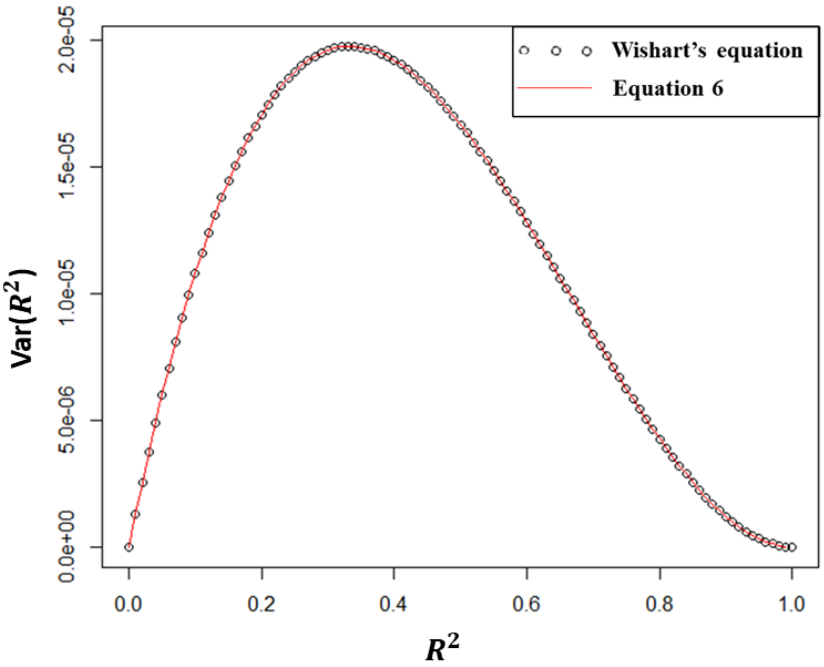

Figure S1: Wishart's equation and Equation 6 provide identical results for any R^2 values ranging from 0 to 1 using sample size 25000.

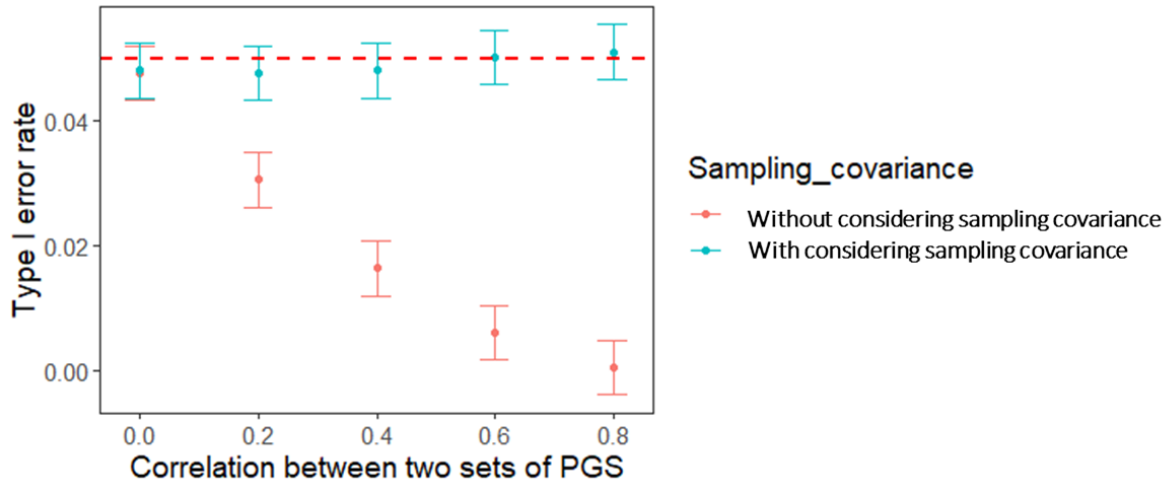

Figure S2: Incorrect type I error rate for testing the difference between two R^2 values (r^2_{y,x_1} vs. r^2_{y,x_2}) when ignoring the correlation between two sets of PGS (r_{x_1,x_2}> 0). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.100 & 0.100 \\ 0.100 & 1 & various \\ 0.100 & various & 1 \end{bmatrix}$, and r^2_{y,x_1} and r^2_{y,x_2} were obtained from models $y = x_1 + e$ and $y = x_2 + e$ in each replicate (the type I error rate was obtained over 10,000 replicates). In each replicate, chi-squared test is used, i.e. $\chi^2_1 = \frac{d^2}{\sigma^2_d}$, where d = $r_{y,x_1} - r_{y,x_2}$ and $\sigma^2_d = \sigma^2_{r^2_{y,x_1}} + \sigma^2_{r^2_{y,x_2}}$ when ignoring the covariance term, and $\sigma^2_d = \sigma^2_{r^2_{y,x_1}} + \sigma^2_{r^2_{y,x_2}} - 2cov(r^2_{y,x_1}, r^2_{y,x_2})$ when considering the covariance term. The sample size is 25,000 in each replication. We used a significance level at p value = 0.05 (red dashed line).

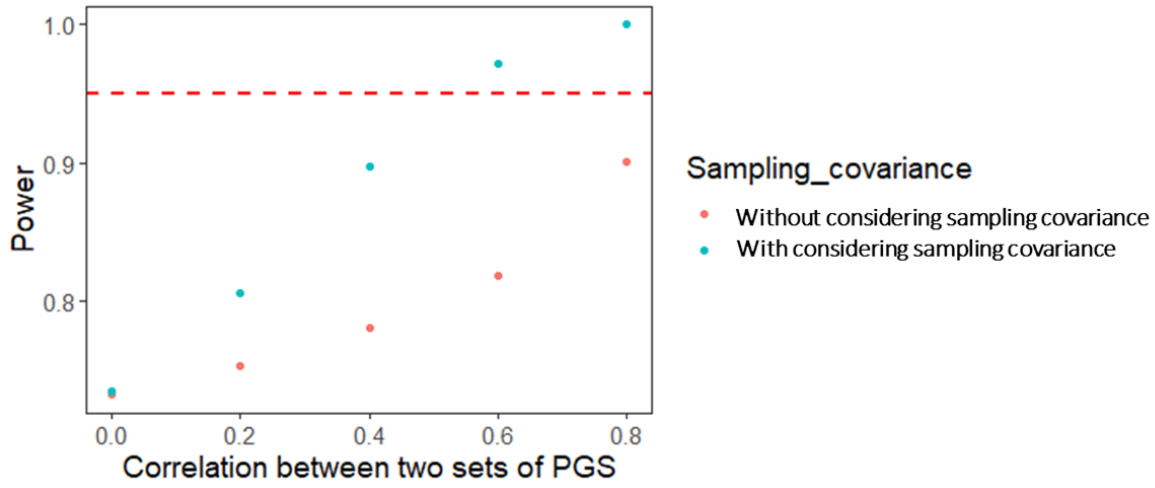

Figure S3: Reduced power for testing the difference between two R^2 values (r^2_{y,x_1} vs. r^2_{y,x_2}) when ignoring the correlation between two sets of PGS (r_{x_1,x_2} > 0). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.120 & 0.100 \\ 0.120 & 1 & various \\ 0.100 & various & 1 \end{bmatrix}$, and r^2_{y,x_1} and r^2_{y,x_2} were obtained from models $y = x_1 + e$ and $y = x_2 + e$ in each replicate (the power was obtained over 10,000 replicates). In each replicate, chi-squared test is used (the same as in Figure S1). The sample size is 25,000 in each replication.

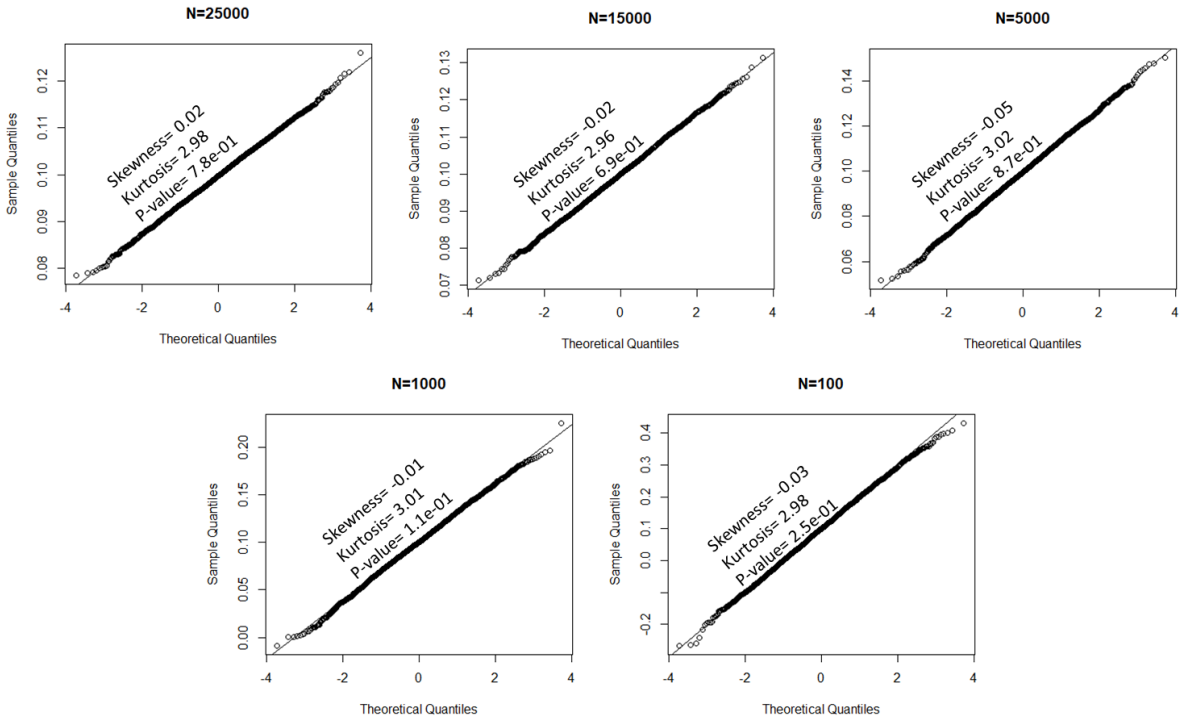

Figure S4: Distribution of regression coefficients are asymptotically normal when correlation between two PGS is 0.10. Simulations of y and x_1 were based on a correlation of $r_{y,x_1} = 0.10$ and the regression coefficient was estimated from a model $y = x_1 + e$ using 5,000 replications. The sample size varies from N=100 to N=25,000. The p value is to test the normality of estimated regression coefficients, using Shapiro-Wilk test, i.e. P < 0.05 means that the regression coefficients are not normally distributed. Skewness and kurtosis are close to 0 and 3 if regression coefficients are normally distributed. For $r_{y,x_1} = 0.10$, the regression coefficients are approximately normal for all the sample sizes considered (N=100 – 25,000).

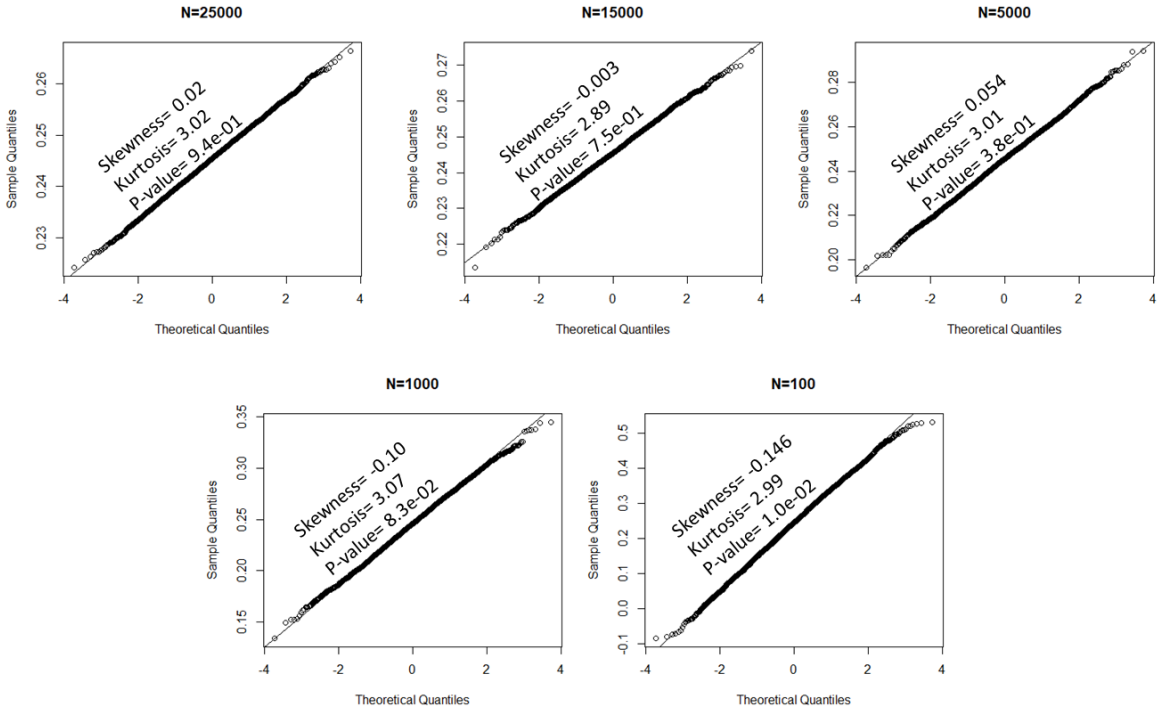

Figure S5: Distribution of regression coefficients are asymptotically normal when correlation between two PGS is 0.25. Simulations of y and x_1 were based on a correlation of $r_{y,x_1} = 0.25$ and the regression coefficient was estimated from a model $y = x_1 + e$ using 5,000 replications. The sample size varies from N=100 to N=25,000. The p value is to test the normality of estimated regression coefficients, using Shapiro-Wilk test, i.e. P < 0.05 means that the regression coefficients are not normally distributed. Skewness and kurtosis are close to 0 and 3 if regression coefficients are normally distributed. For $r_{y,x_1} = 0.25$, the regression coefficients are approximately normal for all the sample sizes investigated except N=100.

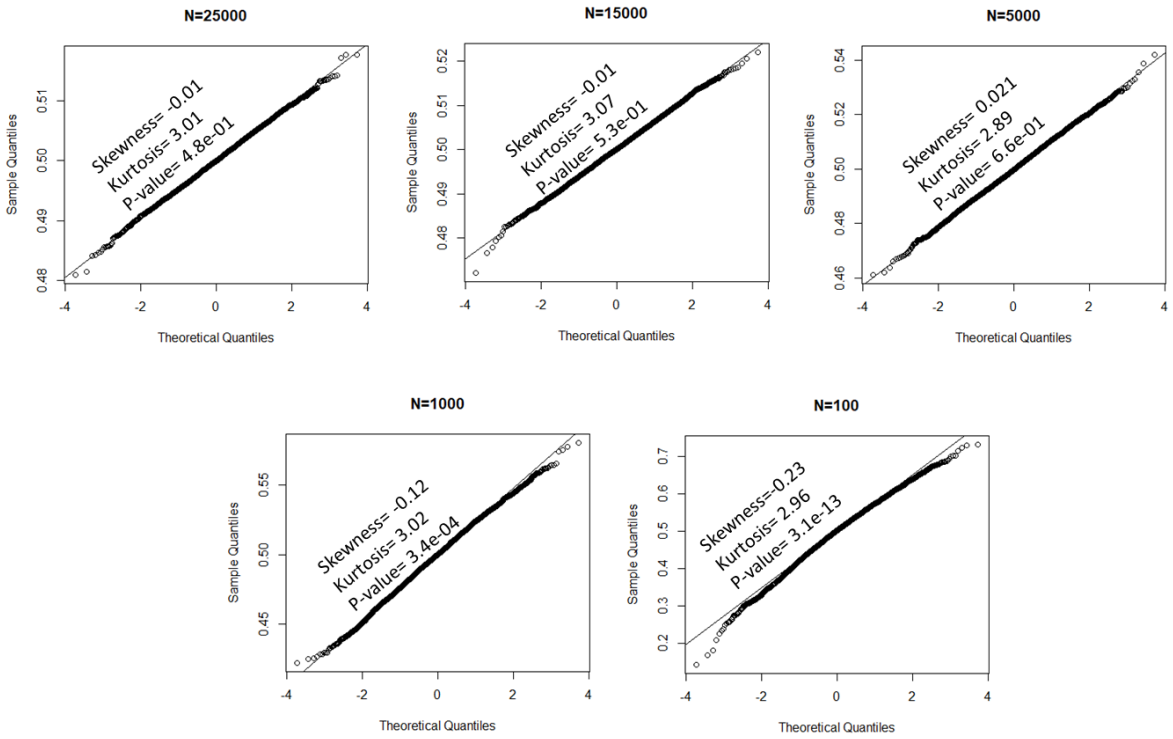

Figure S6: Distribution of regression coefficients are asymptotically normal when correlation between two PGS is 0.50. Simulations of y and x_1 were based on a correlation of $r_{y,x_1} = 0.50$ and the regression coefficient was estimated from a model $y = x_1 + e$ using 5,000 replications. The sample size varies from N=100 to N=25,000. The p value is to test the normality of estimated regression coefficients, using Shapiro-Wilk test, i.e. P < 0.05 means that the regression coefficients are not normally distributed. Skewness and kurtosis are close to 0 and 3 if regression coefficients are normally distributed. For $r_{y,x_1} = 0.50$, the regression coefficients are approximately normal when the sample sizes investigated was N> 5000.

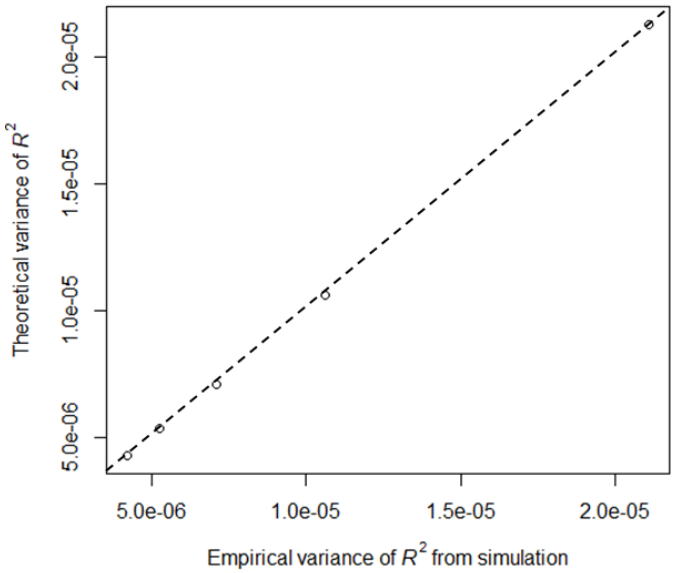

Figure S7: Nearly identical values between the theoretical and empirical variances of R^2 (r^2_{y,x_1}) estimated from 10,000 simulated replicates when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.246 & 0.139 \\ 0.246 & 1 & 0.315 \\ 0.139 & 0.315 & 1 \end{bmatrix}$ and R^2 (r^2_{y,x_1}) was obtained from a model $y = x_1 + e$ in each replicate. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (6). Each data point in the diagonal represents the variance of R^2 with a sample size of 50000, 40000, 30000, 20000 and 10000.

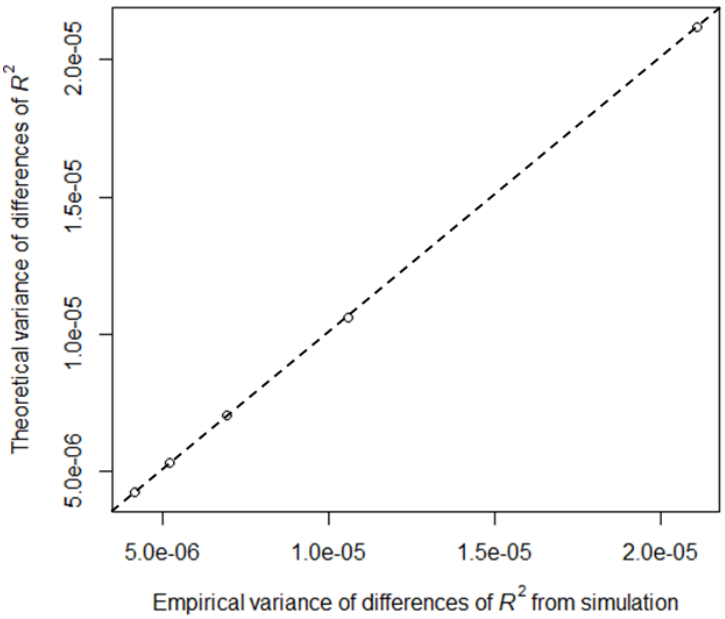

Figure S8: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,x_1} - r^2_{y,x_2}$) estimated from 10,000 simulated replicates when varying sample size.

Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.246 & 0.139 \\ 0.246 & 1 & 0.315 \\ 0.139 & 0.315 & 1 \end{bmatrix}$, and r^2_{y,x_1} and r^2_{y,x_2} were obtained from models $y = x_1 + e$ and $y = x_2 + e$, respectively, to get their difference in each replicate. The empirical variance of $r^2_{y,x_1} - r^2_{y,x_2}$ over 10,000 replicates was estimated. The theoretical variance of $r^2_{y,x_1} - r^2_{y,x_2}$ was obtained from eq. (9). Each data point in the diagonal represents the variance of $r^2_{y,x_1} - r^2_{y,x_2}$ with a sample size of 50000, 40000, 30000, 20000 and 10000.

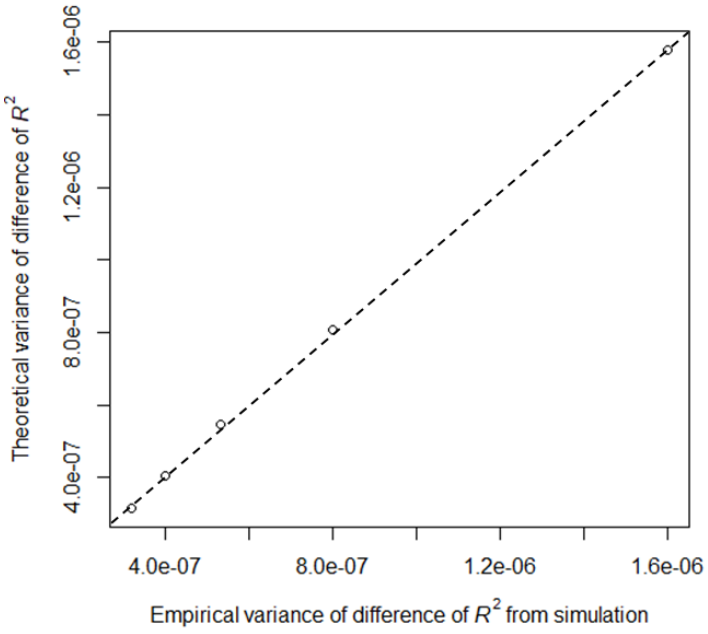

Figure S9: Nearly identical values between the theoretical and empirical variances of R^2 difference of ($r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$) estimated from 10,000 simulated replicates when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.246 & 0.139 \\ 0.246 & 1 & 0.315 \\ 0.139 & 0.315 & 1 \end{bmatrix}$, and $r^2_{y,(x_1,x_2)}$ and r^2_{y,x_1} were obtained from models $y = x_1 + x_2 + e$ and $y = x_1 + e$, respectively, to get their difference in each replicate. The empirical variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ over 10,000 replicates was estimated. The theoretical variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ was obtained from eq. (11). Each data point in the diagonal represents the variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ with a sample size of 50000, 40000, 30000, 20000 and 10000.

.

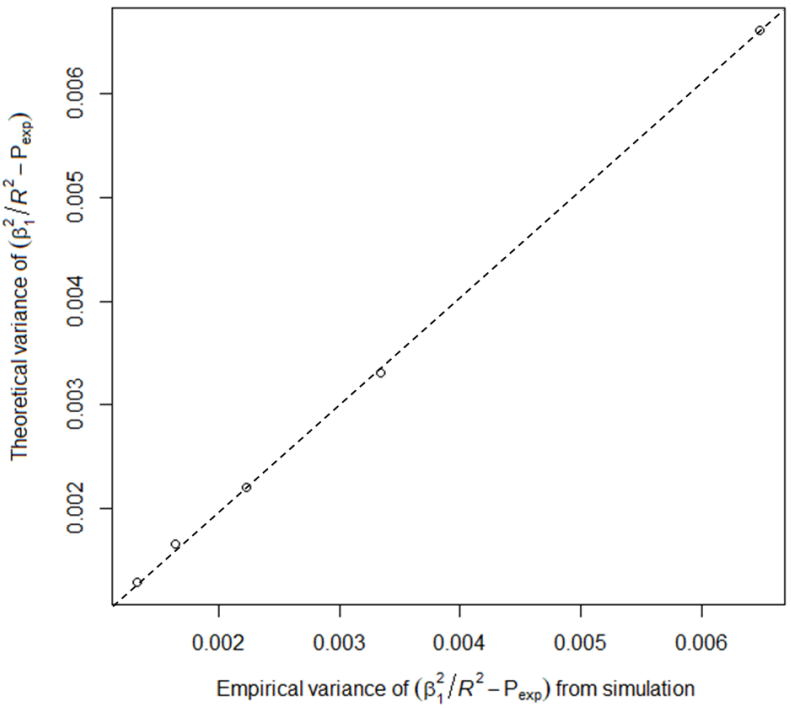

Figure S10: Nearly identical values between the theoretical and empirical variances of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ estimated from 10,000 simulated replicates when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.176 & 0.148 \\ 0.176 & 1 & 0.610 \\ 0.148 & 0.610 & 1 \end{bmatrix}$, and $\hat{\beta}_1^2$ and R^2 were obtained from a multiple regression model $y = x_1 + x_2 + e$ to get the proportion of the coefficient of determination explained by x_1 in each replicate. It was assumed that the expectation is known (p_{exp}= 0.04 was used). The empirical variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ over 10,000 replicates was estimated. The theoretical variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ was obtained from eq. (16). Each data point in the diagonal represents the variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ with a sample size of 50000, 40000, 30000, 20000 and 10000.

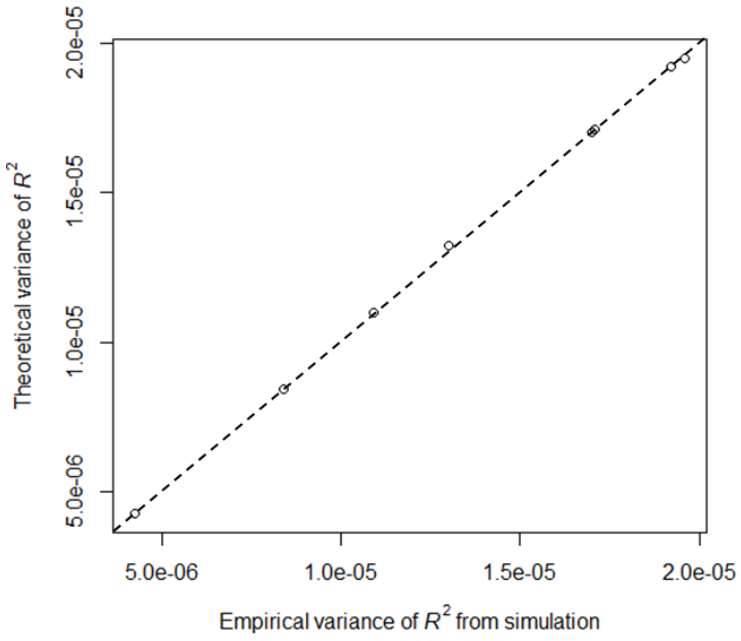

Figure S11: Nearly identical values between the theoretical and empirical variances of R^2 (r^2_{y,x_1}) estimated from 10,000 simulated replicates when varying R^2 value. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.447 \\ various & 1 & 0.800 \\ 0.447 & 0.800 & 1 \end{bmatrix}$ and R^2 (r^2_{y,x_1}) was obtained from a model $y = x_1 + e$ in each replicate. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (6). A sample size of 30,000 was used. Each data point in the diagonal represents the variance of R^2 with r^2_{y,x_1} = 0.80, 0.70, 0.10, 0.60, 0.50, 0.20, 0.40 and 0.30.

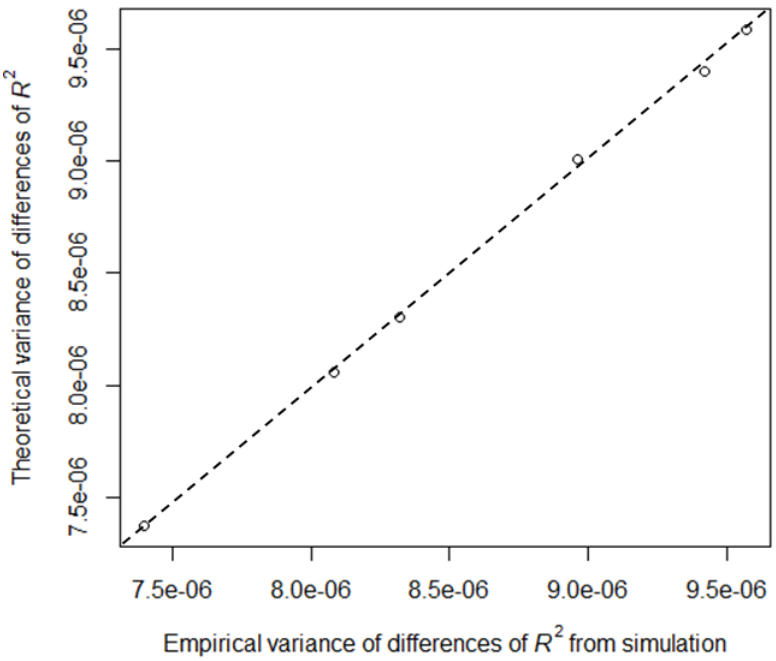

Figure S12: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,x_1} - r^2_{y,x_2}$) estimated from 10,000 simulated replicates when varying R^2 difference.

Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.447 \\ various & 1 & 0.800 \\ 0.447 & 0.800 & 1 \end{bmatrix}$, and r^2_{y,x_1} and r^2_{y,x_2} were obtained from models $y = x_1 + e$ and $y = x_2 + e$, respectively, to get their difference in each replicate. The empirical variance of $r^2_{y,x_1} - r^2_{y,x_2}$ over 10,000 replicates was estimated. The theoretical variance of $r^2_{y,x_1} - r^2_{y,x_2}$ was obtained from eq. (9). A sample size of 30,000 was used. Each data point in the diagonal represents the variance of $r^2_{y,x_1} - r^2_{y,x_2}$ with $r^2_{y,x_1} - r^2_{y,x_2}$ = 0.50, 0.40, 0, 0.30, 0.10 and 0.20.

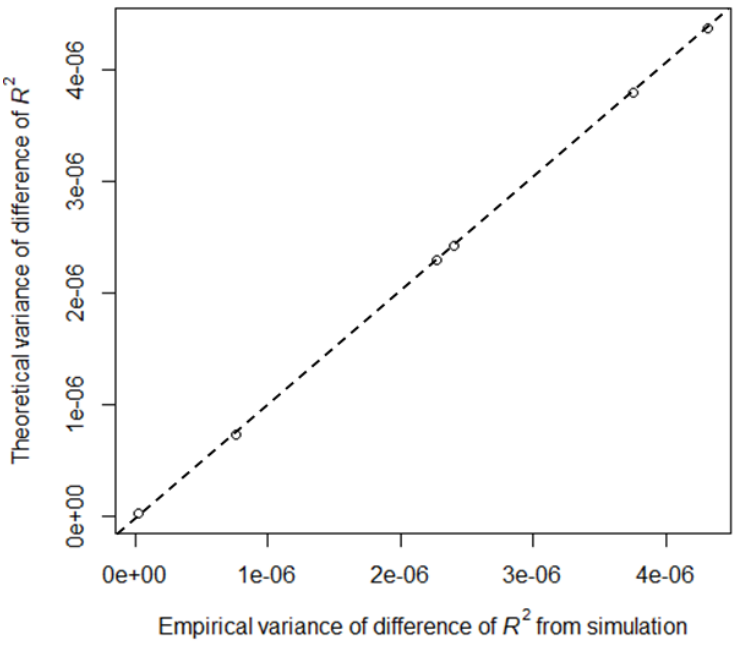

Figure S13: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$) estimated from 10,000 simulated replicates when varying R^2 difference. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.447 \\ various & 1 & 0.800 \\ 0.447 & 0.800 & 1 \end{bmatrix}$, and $r^2_{y,(x_1,x_2)}$ and r^2_{y,x_1} were obtained from models $y = x_1 + x_2 + e$ and $y = x_1 + e$, respectively, to get their difference in each replicate. The empirical variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ over 10,000 replicates was estimated. The theoretical variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ was obtained from eq. (11). A sample size of 30,000 was used. Each data point in the diagonal represents the variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ with $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ = 0.10, 0.20, 0, 0.30, 0.40 and 0.50.

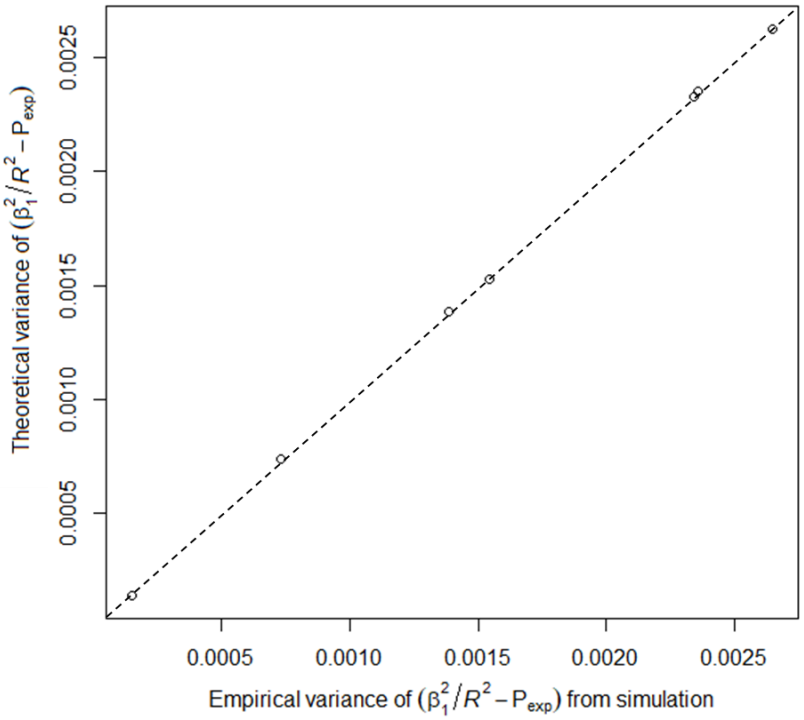

Figure S14. Nearly identical values between the theoretical and empirical variances of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ estimated from 10,000 simulated replicates when varying correlation structure. Simulations of *y*, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.148 \\ various & 1 & 0.610 \\ 0.148 & 0.610 & 1 \end{bmatrix}$, and $\hat{\beta}_1^2$ and R^2 were obtained from a multiple regression model $y = x_1 + x_2 + e$ to get the proportion of the coefficient of determination explained by x_1 in each replicate. It was assumed that the expectation is known (p_{exp} = 0.04 was used). The empirical variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ over 10,000 replicates was estimated. The theoretical variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ was obtained from eq. (16). A sample size of 30,000 was used. Each data point in the diagonal represents the variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ with r_{y,x_1} = 0.10, 0.30, 0.25, 0.05, 0.15, 0.20 and 0.176 (resulting in $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ = -0.026, 1.173, 0.994, 0.130, 0.286, 0.703 and 0.516).

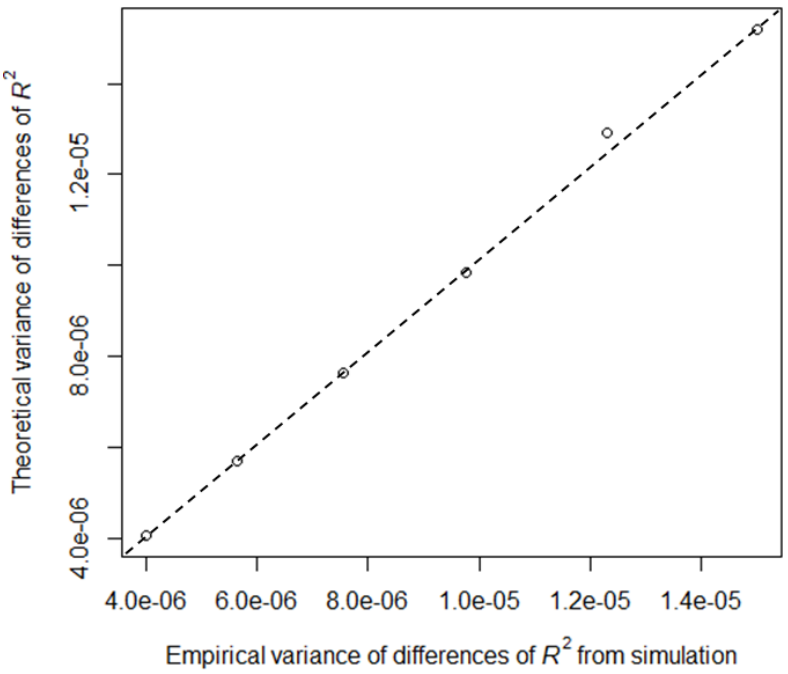

Figure S15: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y_1,x_1} - r^2_{y_2,x_2}$) estimated from 10,000 simulated replicates when using two sets of independent PGS (e.g., Male vs female PGS). Simulations of y_1 and x_1 were based on a correlation structure $\begin{bmatrix} 1 & r_{y_1,x_1} \\ r_{y_1,x_1} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various \\ various & 1 \end{bmatrix}$, simulations of y_2 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y_2,x_2} \\ r_{y_2,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.447 \\ 0.447 & 1 \end{bmatrix}$ and $r^2_{y_1,x_1}$ and $r^2_{y_2,x_2}$ were obtained from models $y_1 = x_1 + e$ and $y_2 = x_2 + e$, respectively, to get their difference in each replicate. The empirical variance of $r^2_{y1,x_1} - r^2_{y2,x_2}$ over 10,000 replicates was estimated. The theoretical variance of $r^2_{y_1,x_1} - r^2_{y_2,x_2}$ was obtained from eq. (14). A sample size of 15,000 and 17,000 were used for 1st and 2nd PGS, respectively. Each data point in the diagonal represents the variance of $r^2_{y_1,x_1} - r^2_{y_2,x_2}$ with $r^2_{y_1,x_1} - r^2_{y_2,x_2}$ = 0, 0.02, 0.04, 0.06, 0.08 and 0.10.

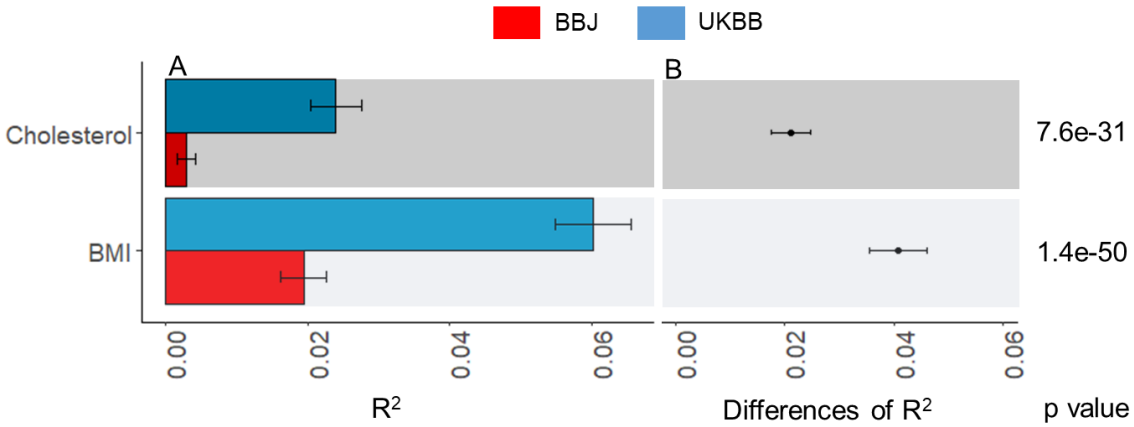

Figure S16: The predictive ability of (R^2) for male and female, when predicting European male and female separately using UKBB and BBJ discovery samples.
Panel A: The main bars represent R^2 values and error bars correspond 95% confidence intervals.
Panel B: Dot points represent the differences of R^2 values between male and female PGS models, and error bars indicate 95% confidence intervals of the difference.
95% confidence interval for the differences of R^2 between two independent sets of PGS (male and female) was estimated from eq. (15).

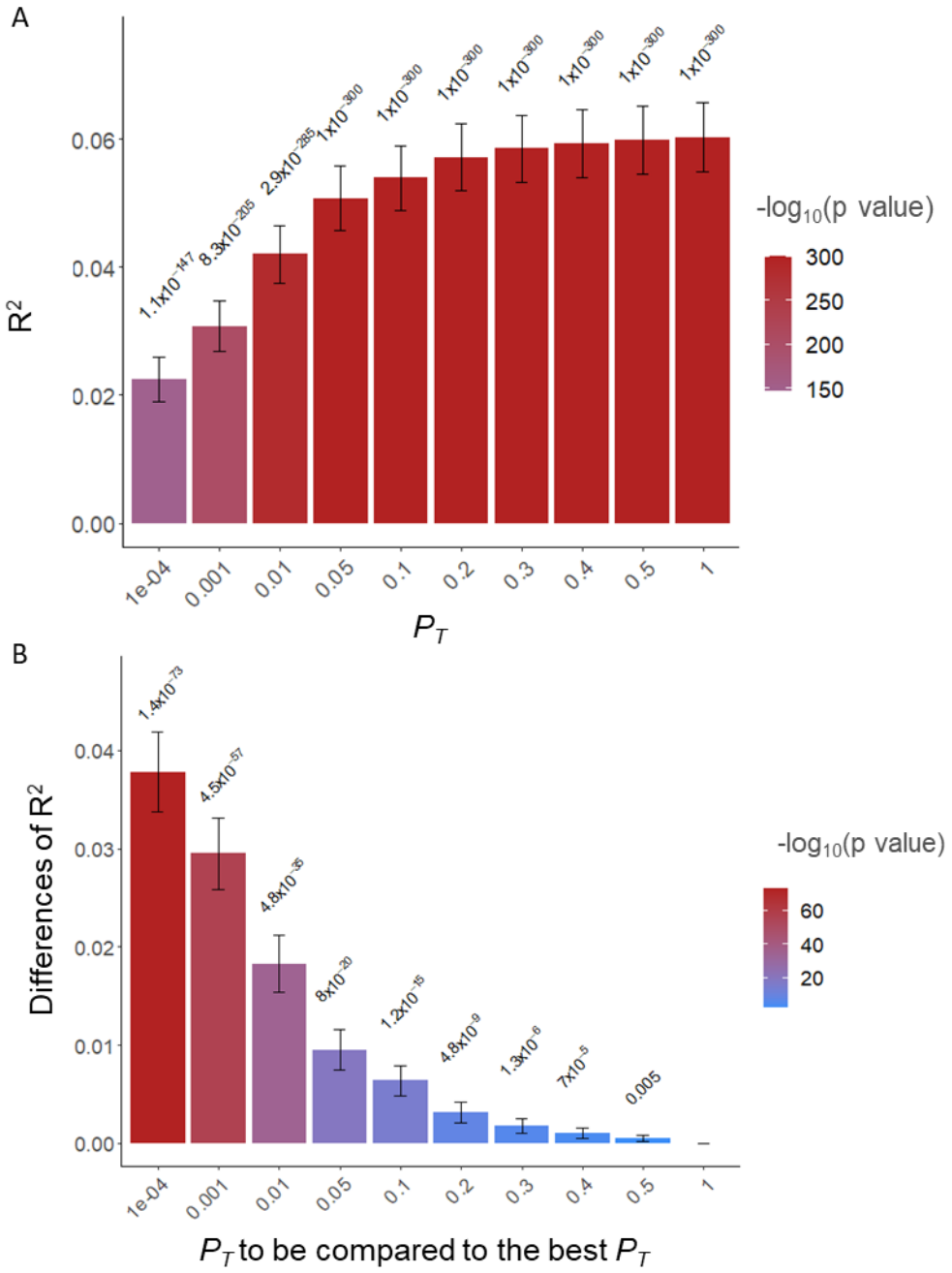

Figure S17: The predictive ability (R^2) of PGS estimated based on SNPs below the p_T when predicting BMI in 28,880 European samples using UKBB discovery samples (GWAS summary statistics).

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 threshold and the best-performing threshold and error bars indicate 95% confidence intervals. The values above 95% CIs are p values indicating the significance of the difference between the pairs of R^2 values.

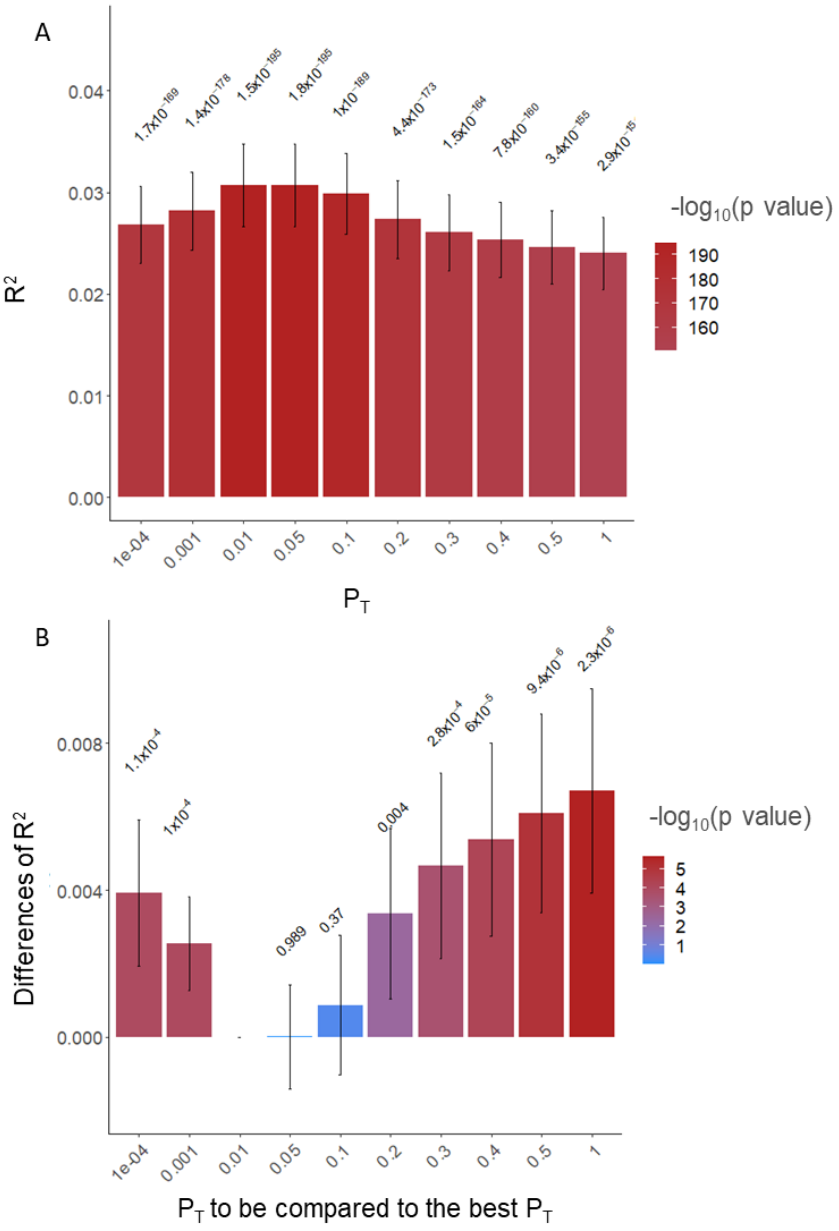

Figure S18: The predictive ability (R^2) of PGS estimated based on SNPs below the p_T when predicting cholesterol in 28,880 European samples using UKBB discovery samples (GWAS summary statistics).

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 threshold and the best-performing threshold and error bars indicate 95% confidence intervals. The values above 95% CIs are p values indicating the significance of the difference between the pairs of R^2 values.

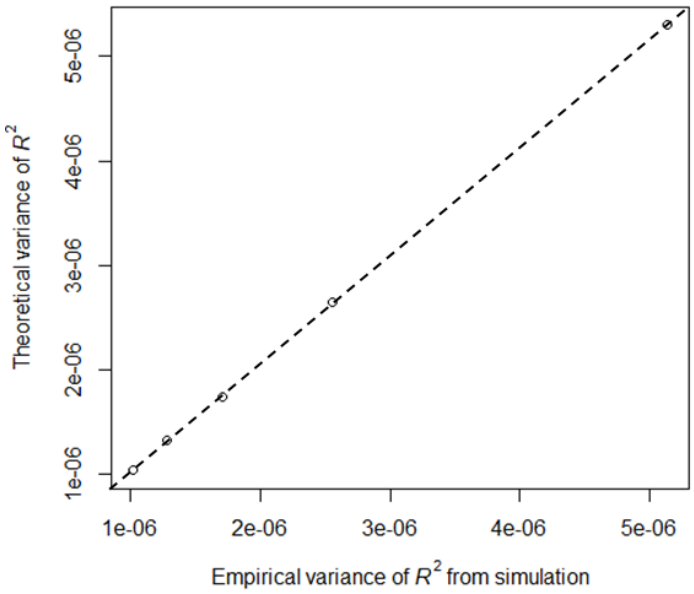

Figure S19: Nearly identical values between the theoretical and empirical variances of R^2 (r^2_{y,x_1}) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.246 & 0.139 \\ 0.246 & 1 & 0.315 \\ 0.139 & 0.315 & 1 \end{bmatrix}$ and R^2 (r^2_{y,x_1}) was obtained from a model $y = x_1 + e$ in each replicate. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (6). Each data point in the diagonal represents the variance of R^2 with a sample size of 50000, 40000, 30000, 20000 and 10000.

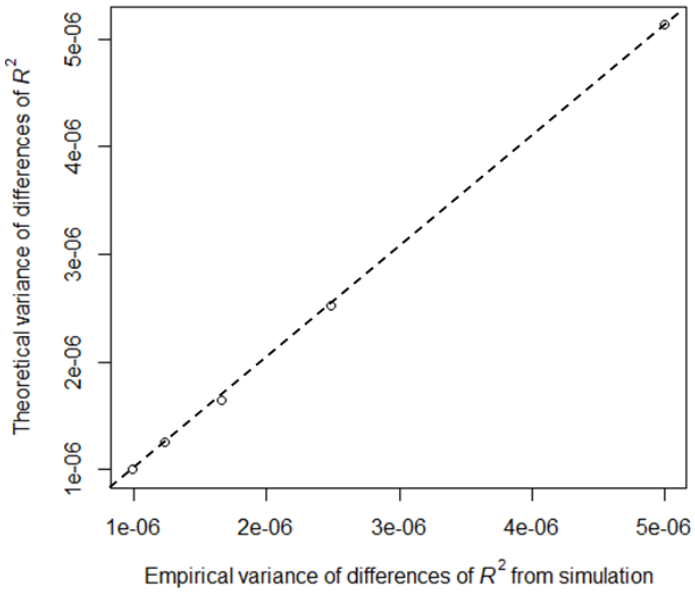

Figure S20: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,x_1} - r^2_{y,x_2}$) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.246 & 0.139 \\ 0.246 & 1 & 0.315 \\ 0.139 & 0.315 & 1 \end{bmatrix}$, and r^2_{y,x_1} and r^2_{y,x_2} were obtained from models $y = x_1 + e$ and $y = x_2 + e$, respectively, to get their difference in each replicate. The empirical variance of $r^2_{y,x_1} - r^2_{y,x_2}$ over 10,000 replicates was estimated. The theoretical variance of $r^2_{y,x_1} - r^2_{y,x_2}$ was obtained from eq. (9). Each data point in the diagonal represents the variance of $r^2_{y,x_1} - r^2_{y,x_2}$ with a sample size of 50000, 40000, 30000, 20000 and 10000.

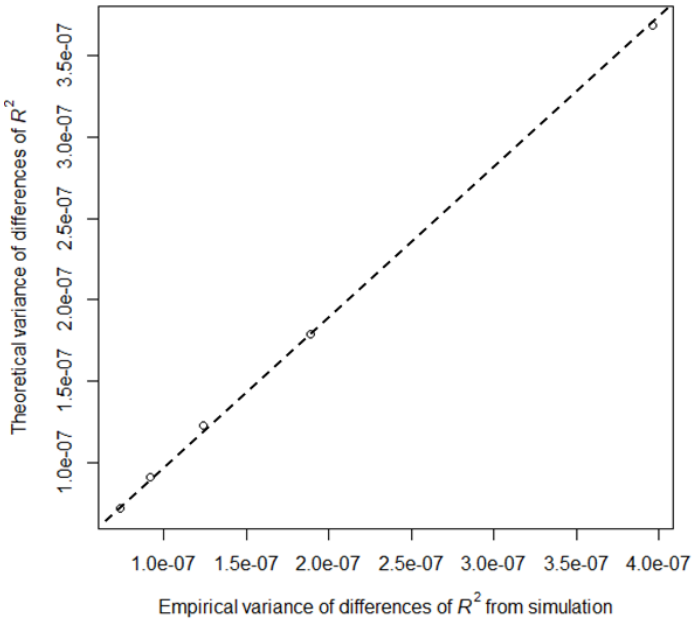

Figure S21: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.246 & 0.139 \\ 0.246 & 1 & 0.315 \\ 0.139 & 0.315 & 1 \end{bmatrix}$, and $r^2_{y,(x_1,x_2)}$ and r^2_{y,x_1} were obtained from models $y = x_1 + x_2 + e$ and $y = x_1 + e$, respectively, to get their difference in each replicate. The empirical variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ over 10,000 replicates was estimated. The theoretical variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ was obtained from eq. (11). Each data point in the diagonal represents the variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ with a sample size of 50000, 40000, 30000, 20000 and 10000.

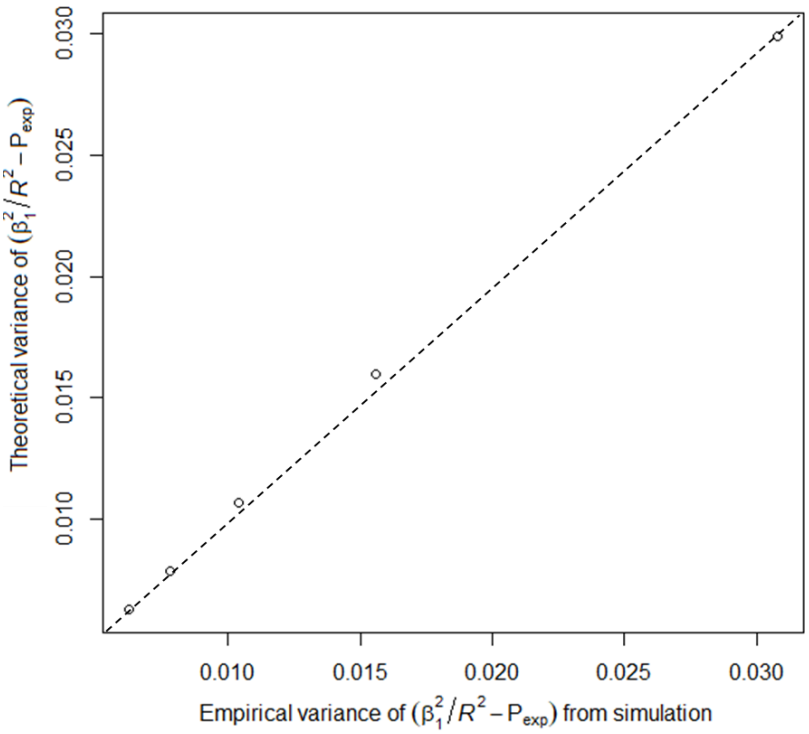

Figure 22: Nearly identical values between the theoretical and empirical variances of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.176 & 0.148 \\ 0.176 & 1 & 0.610 \\ 0.148 & 0.610 & 1 \end{bmatrix}$, and $\hat{\beta}_1^2$ and R^2 were obtained from a multiple regression model $y = x_1 + x_2 + e$ to get the proportion of the coefficient of determination explained by x_1 in each replicate. It was assumed that the expectation is known (p_{exp} = 0.04 was used). The empirical variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ over 10,000 replicates was estimated. The theoretical variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ was obtained from eq. (16). Each data point in the diagonal represents the variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ with a sample size of 50000, 40000, 30000, 20000 and 10000.

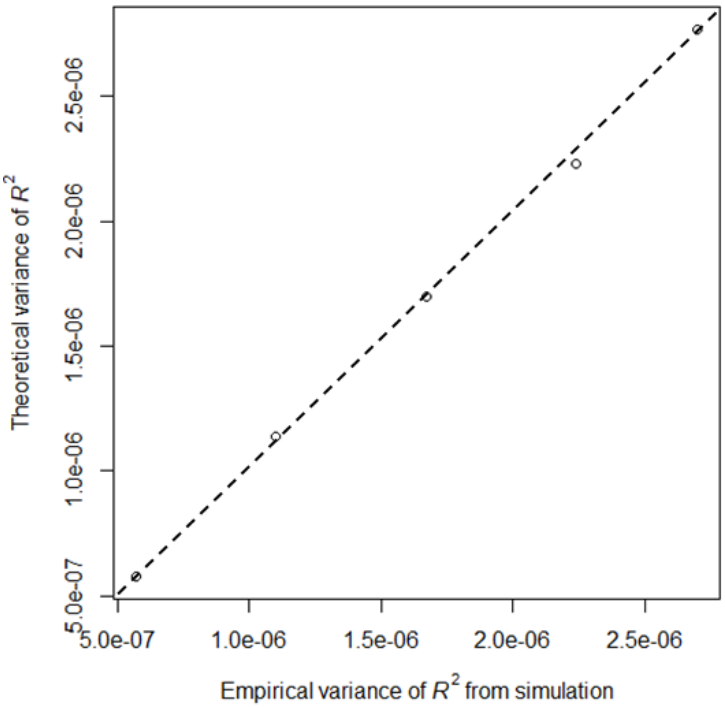

Figure S23: Nearly identical values between the theoretical and empirical variances of R^2 (r^2_{y,x_1}) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying R^2 value. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.141 \\ various & 1 & 0.800 \\ 0.141 & 0.8 & 1 \end{bmatrix}$ and $R^2(r^2_{y,x_1})$ was obtained from a model $y = x_1 + e$ in each replicate. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (6). A sample size of 30,000 was used. Each data point in the diagonal represents the variance of R^2 with r^2_{y,x_1} = 0.02, 0.04, 0.06, 0.08, and 0.10.

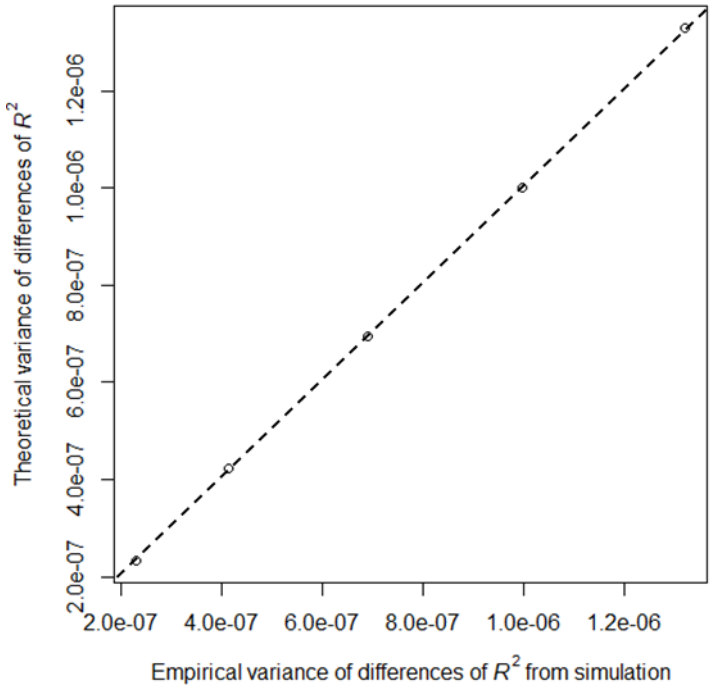

Figure S24: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,x_1} - r^2_{y,x_2}$) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying R^2 difference. Simulations of *y*, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.141 \\ various & 1 & 0.800 \\ 0.141 & 0.8 & 1 \end{bmatrix}$, and r^2_{y,x_1} and r^2_{y,x_2} were obtained from models $y = x_1 + e$ and $y = x_2 + e$, respectively, to get their difference in each replicate. The empirical variance of $r^2_{y,x_1} - r^2_{y,x_2}$ over 10,000 replicates was estimated. The theoretical variance of $r^2_{y,x_1} - r^2_{y,x_2}$ was obtained from eq. (9). A sample size of 30,000 was used. Each data point in the diagonal represents the variance of $r^2_{y,x_1} - r^2_{y,x_2}$ with $r^2_{y,x_1} - r^2_{y,x_2}$ = 0, 0.02, 0.04, 0.0.06, and 0.08.

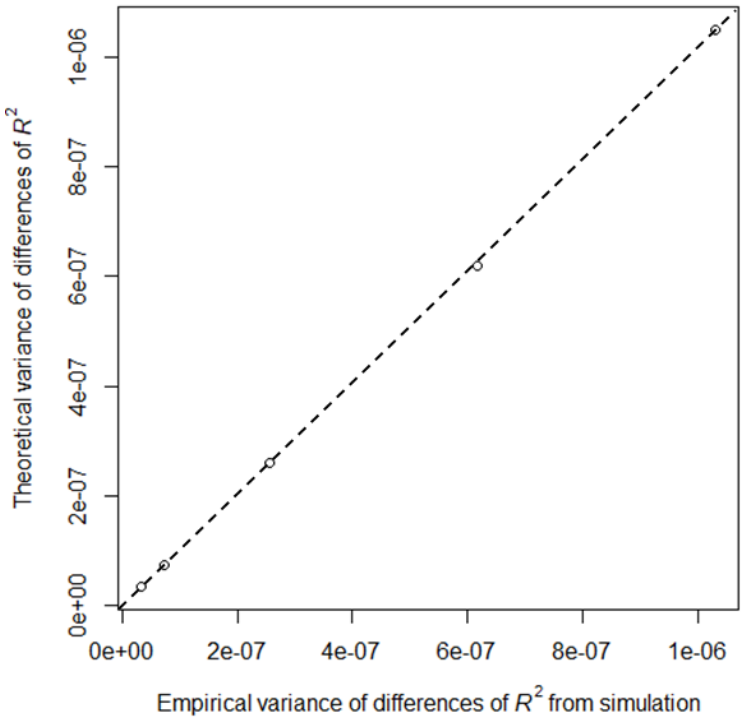

Figure S25: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying R^2 difference. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.141 \\ various & 1 & 0.800 \\ 0.141 & 0.8 & 1 \end{bmatrix}$, and $r^2_{y,(x_1,x_2)}$ and r^2_{y,x_1} were obtained from models $y = x_1 + x_2 + e$ and $y = x_1 + e$, respectively, to get their difference in each replicate. The empirical variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ over 10,000 replicates was estimated. The theoretical variance of$r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ was obtained from eq. (11). A sample size of 30,000 was used. Each data point in the diagonal represents the variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ with $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ = 0.02, 0.00, 0.04, 0.0.06, and 0.08.

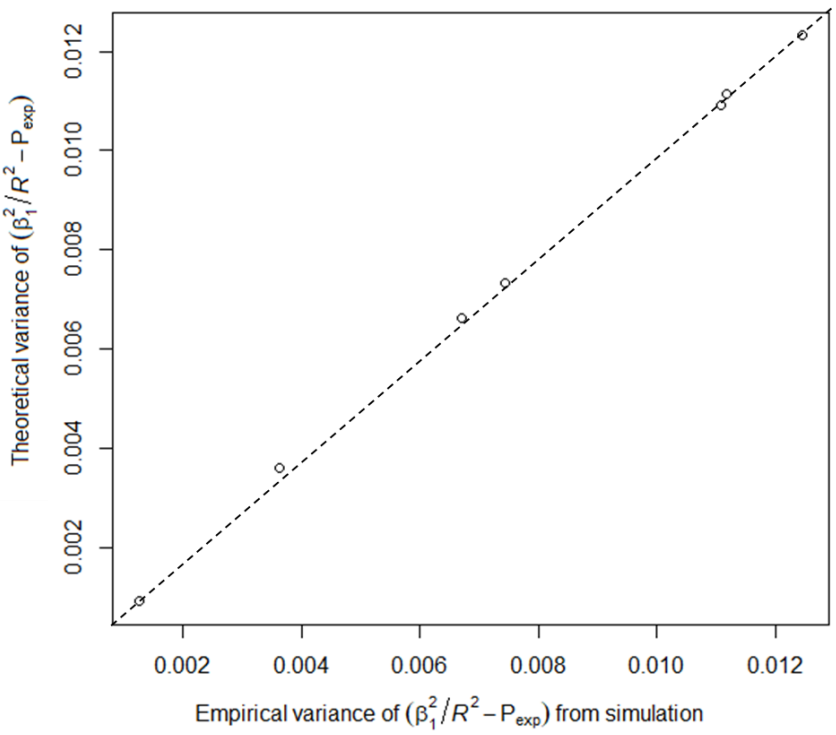

Figure S26: Nearly identical values between the theoretical and empirical variances of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying correlation structure. Simulations of *y*, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.148 \\ various & 1 & 0.610 \\ 0.148 & 0.610 & 1 \end{bmatrix}$, and $\hat{\beta}_1^2$ and R^2 were obtained from a multiple regression model $y = x_1 + x_2 + e$ to get the proportion of the coefficient of determination explained by x_1 in each replicate. It was assumed that the expectation is known (p_{exp} = 0.04 was used). The empirical variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ over 10,000 replicates was estimated. The theoretical variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ was obtained from eq. (16). A sample size of 30,000 was used. Each data point in the diagonal represents the variance $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ with r_{y,x_1} = 0.10, 0.30, 0.25, 0.05, 0.15, 0.20 and 0.176 (resulting in $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ = -0.017, 1.171, 0.993, 0.137, 0.294, 0.702, 0.517 and 0.605).

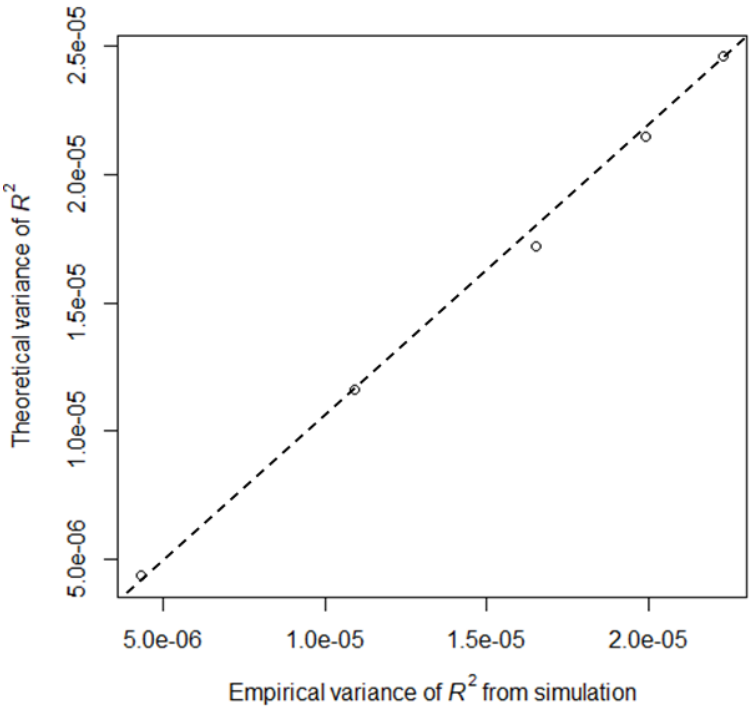

Figure S27: Nearly identical values between the theoretical and empirical variances of R^2 (r^2_{y,x_1}) estimated from 10,000 simulated replicates of ascertained case-control (10000 cases and 10000 controls) assuming 5% disease prevalence and 20000 individuals. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.141 \\ various & 1 & 0.800 \\ 0.141 & 0.8 & 1 \end{bmatrix}$ and R^2 (r^2_{y,x_1}) was obtained from a model $y = x_1 + e$ in each replicate. Following the correlation structure and disease prevalence, we simulated 200,000 dependent and explanatory variables and randomly selected 10000 cases and 10000 controls. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (6). Each data point in the diagonal represents the variance of R^2 with r^2_{y,x_1} = 0.02, 0.04, 0.06, 0.08, and 0.10.

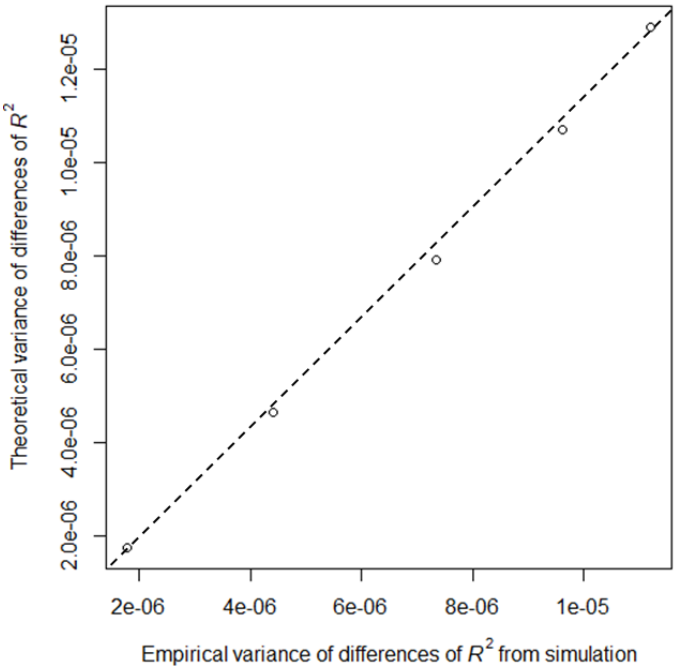

Figure S28: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,x_1} - r^2_{y,x_2}$) estimated from 10,000 simulated replicates of ascertained case-control (10000 cases and 10000 controls) assuming 5% disease prevalence and 20000 individuals. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.141 \\ various & 1 & 0.800 \\ 0.141 & 0.8 & 1 \end{bmatrix}$ and r^2_{y,x_1} and r^2_{y,x_2} were obtained from models $y = x_1 + e$ and $y = x_2 + e$, respectively, to get their difference in each replicate. . Following the correlation structure and disease prevalence, we simulated 200,000 dependent and explanatory variables and randomly selected 10000 cases and 10000 controls. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (9). Each data point in the diagonal represents the variance of $r^2_{y,x_1} - r^2_{y,x_2}$ with $r^2_{y,x_1} - r^2_{y,x_2}$ = 0, 0.02, 0.04, 0.06, and 0.08.

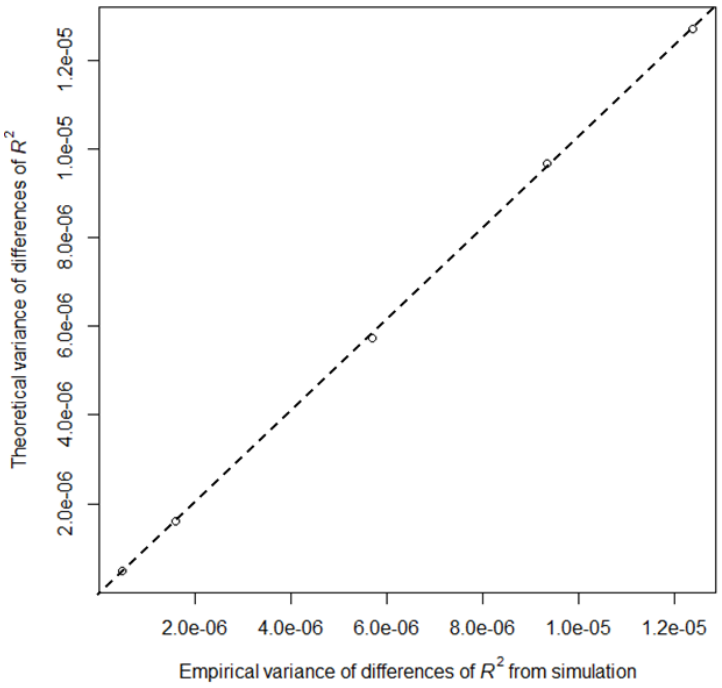

Figure S29: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$) estimated from 10,000 simulated replicates of ascertained case-control (10000 cases and 10000 controls) assuming 5% disease prevalence and 20000 individuals.

Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.141 \\ various & 1 & 0.800 \\ 0.141 & 0.8 & 1 \end{bmatrix}$ and $r^2_{y,(x_1,x_2)}$ and r^2_{y,x_1} were obtained from models $y = x_1 + x_2 + e$ and $y = x_1 + e$, respectively, to get their difference in each replicate. Following the correlation structure and disease prevalence, we simulated 200,000 dependent and explanatory variables and randomly selected 10000 cases and 10000 controls. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (11). Each data point in the diagonal represents the variance of $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ with $r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$ = 0, 0.04, 0.08, 0.12, and 0.16.

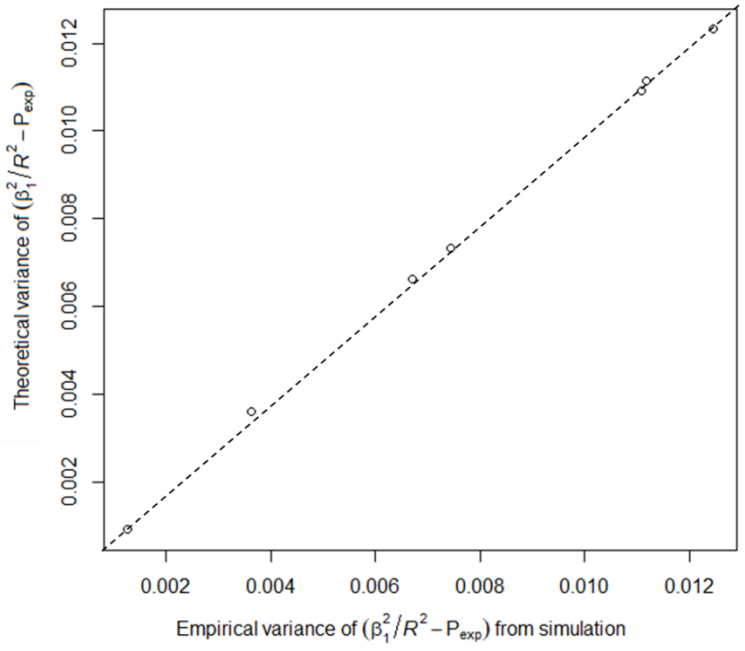

Figure S30: Nearly identical values between the theoretical and empirical variances of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ estimated from 10,000 simulated replicates of ascertained case-control (10000 cases and 10000 controls) assuming 5% disease prevalence and 20000 individuals. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.148 \\ various & 1 & 0.610 \\ 0.148 & 0.610 & 1 \end{bmatrix}$, and $\hat{\beta}_1^2$ and R^2 were obtained from a multiple regression model $y = x_1 + x_2 + e$ to get the proportion of the coefficient of determination explained by x_1 in each replicate. It was assumed that the expectation is known (p_{exp} = 0.04 was used). Following the correlation structure and disease prevalence, we simulated 200,000 dependent and explanatory variables and randomly selected 10000 cases and 10000 controls. The empirical variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ over 10,000 replicates was estimated. The theoretical variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ was obtained from eq. (17). Each data point in the diagonal represents the variance of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ with r_{y,x_1} = 0.10, 0.30, 0.25, 0.05, 0.15, 0.20 and 0.176 (resulting in $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ = -0.026, 1.172, 0.995, 0.127, 0.0.288, 0.703 and 0.514).

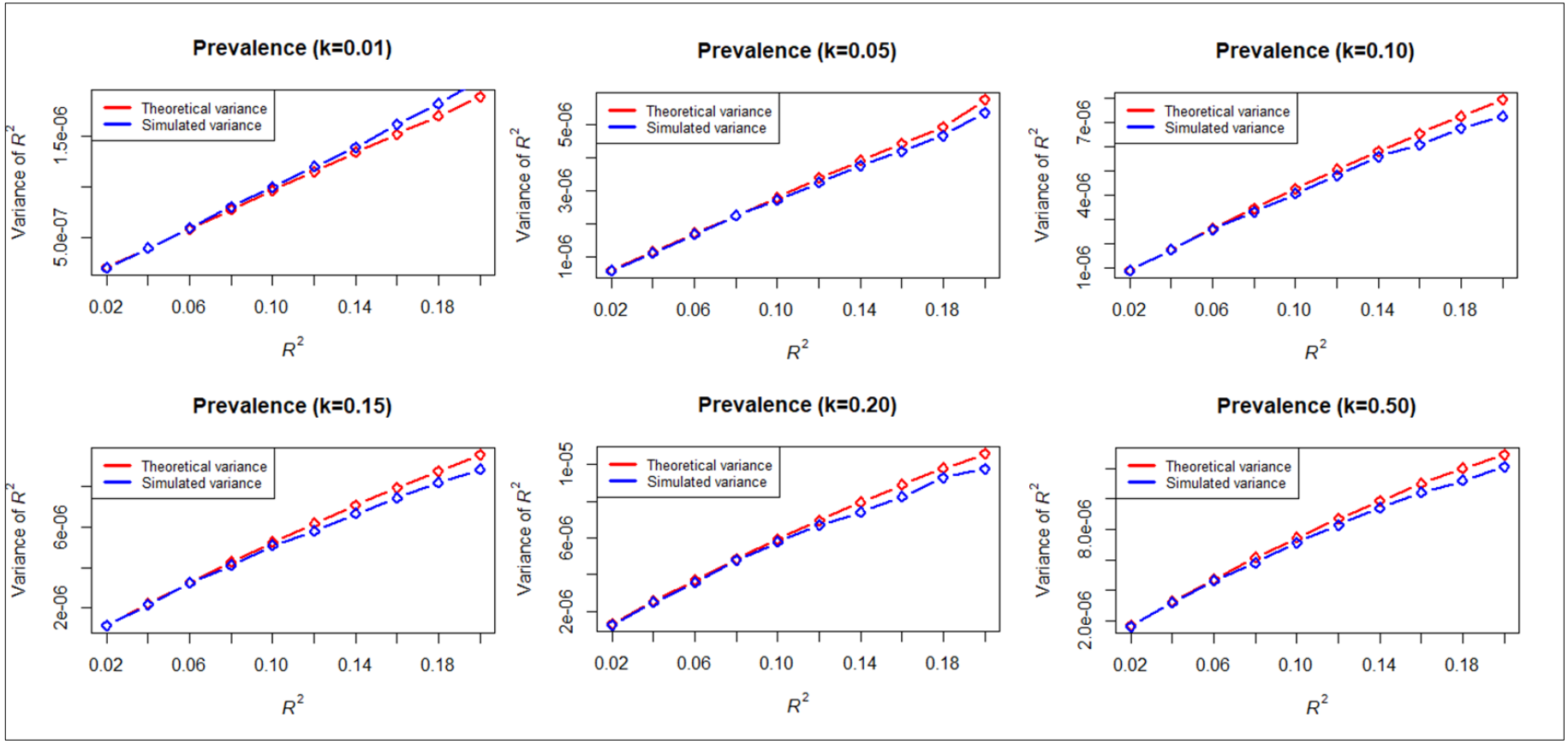

Figure S31: The empirical and theoretical variances diverge when R^2 values are more than 0.1 for binary responses, noting that $R^2 > 0.1$ is not frequently observed (see Supplemental Table 2). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.141 \\ various & 1 & 0.800 \\ 0.141 & 0.8 & 1 \end{bmatrix}$ and R^2 (r^2_{y,x_1}) was obtained from a model $y = x_1 + e$ in each replicate. Following the correlation structure and disease prevalence, we simulated 30,000 dependent and explanatory variables. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (6). Each data point represents the variance of R^2 ranged from 0.02 to 0.2.

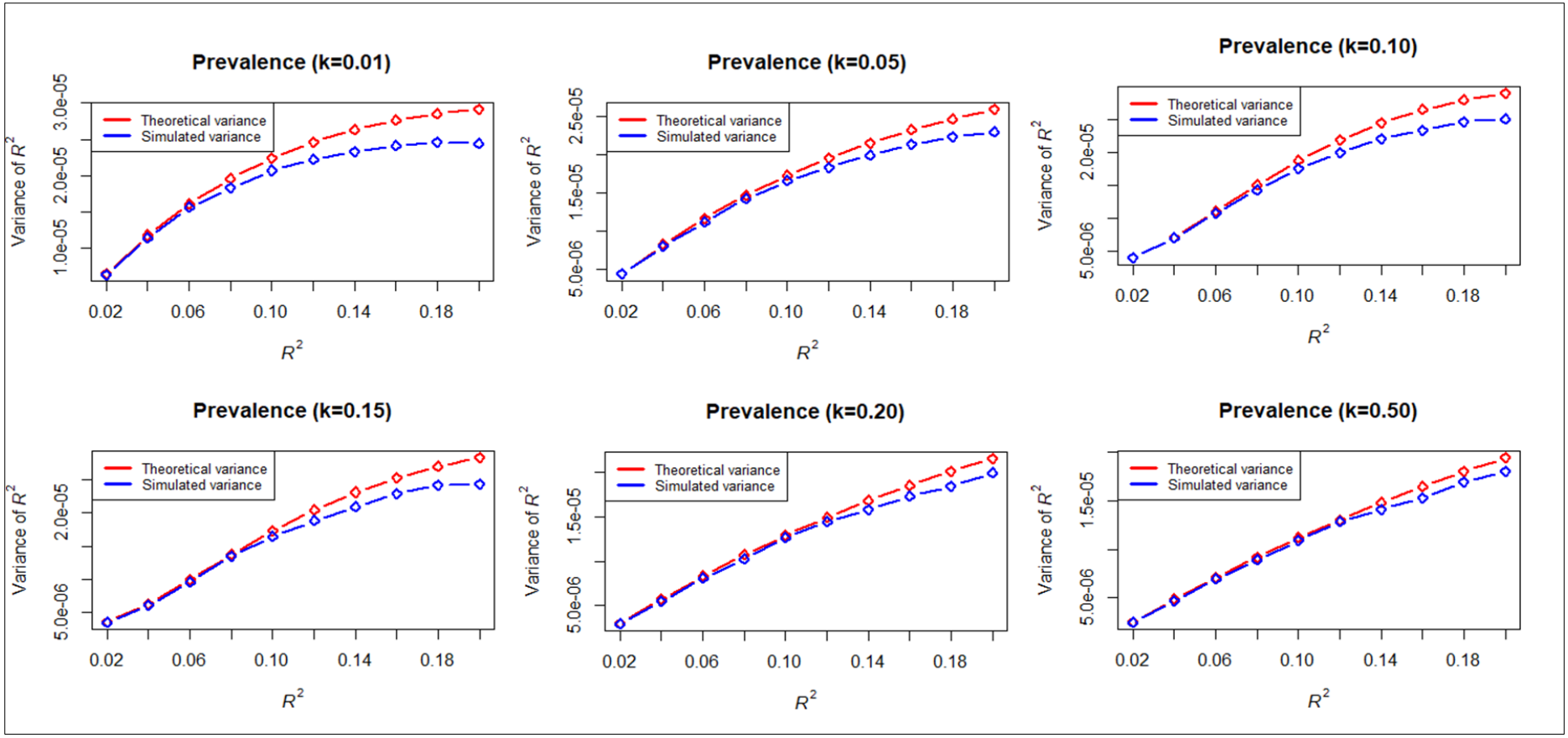

Figure S32: The empirical and theoretical variances become disagreed when R^2 values are more than 0.1 for ascertained case-control samples in the reference dataset (10000 cases and 10000 controls), noting that R^2 > 0.1 is not frequently observed (see Supplemental Table 2). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.141 \\ various & 1 & 0.800 \\ 0.141 & 0.8 & 1 \end{bmatrix}$ and R^2 (r^2_{y,x_1}) was obtained from a model $y = x_1 + e$ in each replicate. Following the correlation structure and disease prevalence, we simulated 200,000 dependent and explanatory variables and randomly selected 10000 cases and 10000 controls. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (6). Each data point represents the variance of R^2 ranged from 0.02 to 0.2.

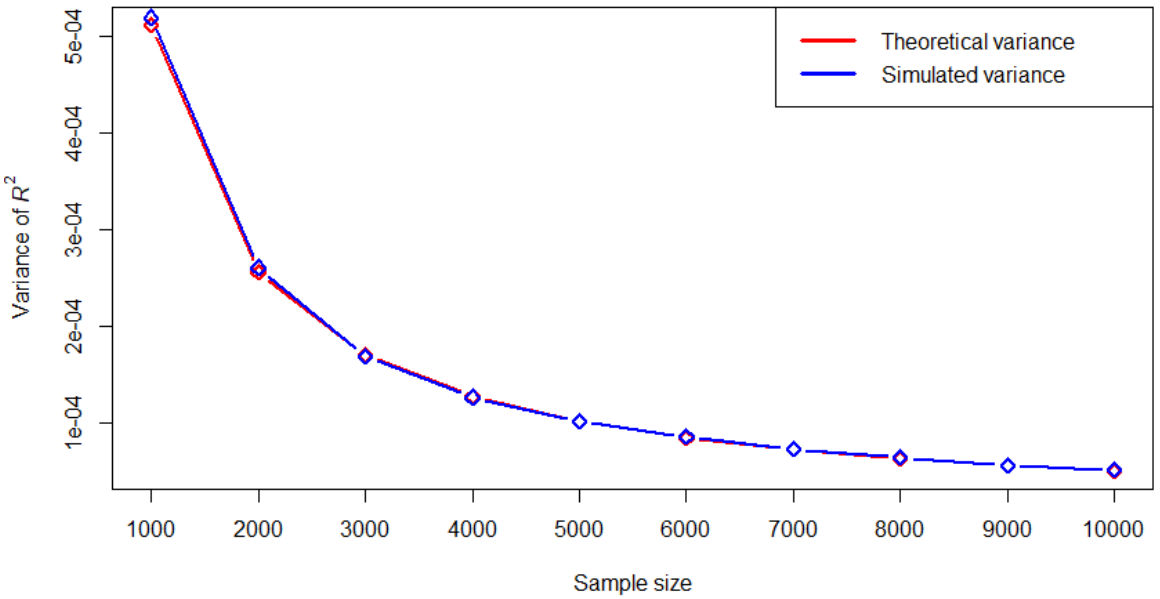

Figure S33: The empirical and theoretical variances agree even with sample size 2000 for quantitative phenotypes. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.44 & 0.31 \\ 0.44 & 1 & 0.800 \\ 0.31 & 0.8 & 1 \end{bmatrix}$ and R^2 (r^2_{y,x_1}) was obtained from a model $y = x_1 + e$ in each replicate. Following the correlation structure and disease prevalence, we simulated dependent and explanatory variables. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (6). Each data point represents the variance of R^2 for different sample size.

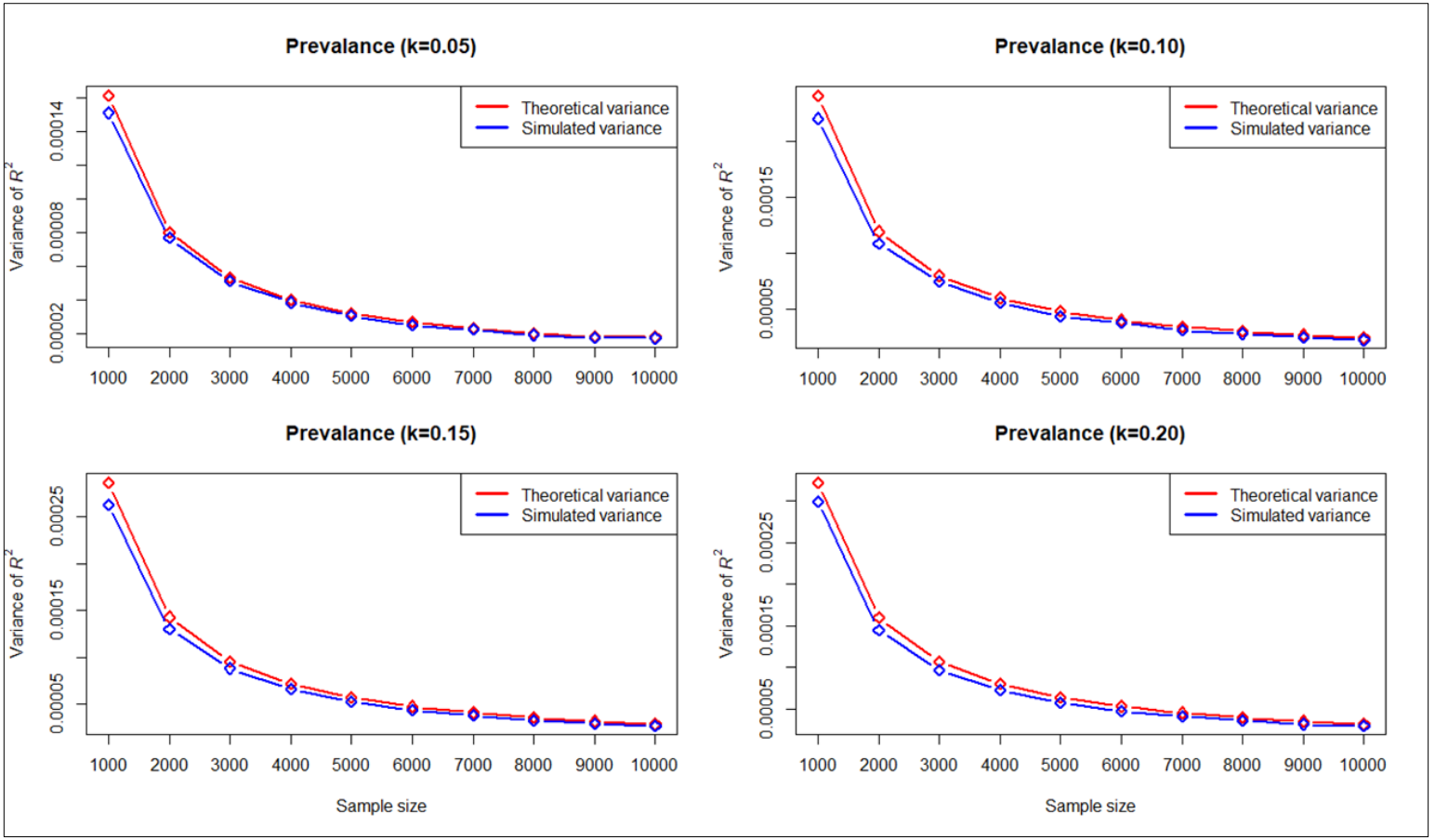

Figure S34: The empirical and theoretical variances become disagreed when sample size is < 5000 for binary responses under scenario of different prevalence rate (k). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.44 & 0.31 \\ 0.44 & 1 & 0.800 \\ 0.31 & 0.8 & 1 \end{bmatrix}$ and R^2 (r^2_{y,x_1}) was obtained from a model $y = x_1 + e$ in each replicate. Following the correlation structure and disease prevalence, we simulated dependent and explanatory variables. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (6). Each data point represents the variance of R^2 for different sample size.

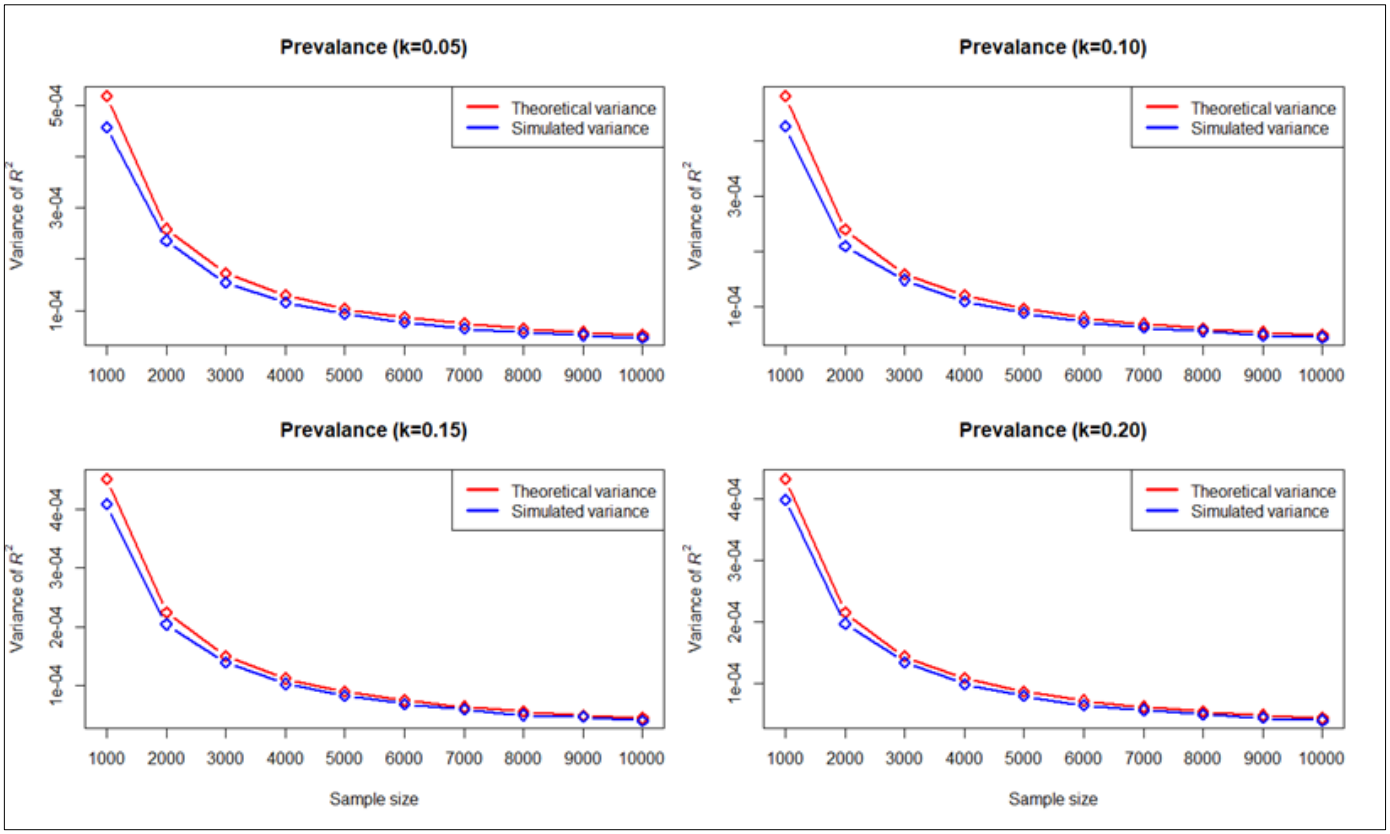

Figure S35: The empirical and theoretical variances become disagreed when sample size is < 5000 for ascertained case-control samples in the reference dataset (50% cases and 50% controls) under scenario of different prevalence rate (k). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.44 & 0.31 \\ 0.44 & 1 & 0.800 \\ 0.31 & 0.8 & 1 \end{bmatrix}$ and R^2 (r^2_{y,x_1}) was obtained from a model $y = x_1 + e$ in each replicate. Following the correlation structure and disease prevalence, we simulated 100,000 dependent and explanatory variables and randomly selected cases and controls. The empirical variance of R^2 over 10,000 replicates was estimated. The theoretical variance of R^2 was obtained from eq. (6). Each data point represents the variance of R^2 for different sample size.

Supplemental tables

P value Threshold	BMI		Cholesterol	
	No of SNPs (UKBB)	No of SNPs (BBJ)	No of SNPs (UKBB)	No of SNPs (BBJ)
1	4113630	4113630	4113630	4113630
0.5	2539432	2365077	2254467	2143406
0.4	2199702	1996741	1864917	1746560
0.3	1841948	1610857	1468402	1346257
0.2	1442727	1201525	1059630	936383
0.1	976865	746948	618466	508651
5e-02	675502	475337	376346	280526
1e-02	318902	128704	140757	76284
1e-03	134943	57272	54829	19216
1e-04	67528	24320	30741	8636

Table S1: Number of SNPs across different p value thresholds for BMI and cholesterol for UKBB and BBJ

Disease	**Prevalence in discovery GWAS (*n*)**	**Prevalence in validation dataset**	**AUC (95% CI) in validation dataset**	**Predictive ability (R^2)**
Coronary Artery disease (CAD)	60,801 cases and 123,504 controls (32.9%)[3]	3,963 cases and 116,317 controls (3.4%)	0.81 (0.80–0.81)	0.040
Atrial fibrillation	17,931 cases and 115,142 controls (13.4%)[4]	2,024 cases and 118,256 controls (1.7%)	0.77 (0.76–0.78)	0.016
Type 2 diabetes	6,676 cases and 132,532 controls (16.7%)[5]	2,785 cases and 117,495 controls (2.4%)	0.72 (0.72–0.73)	0.012
Inflammatory bowel disease	2,882 cases and 21,770 controls (37.2)[6]	1,360 cases and 118,920 controls (1.1%)	0.63 (0.62–0.65)	0.003
Breast cancer	122,977 cases and 105,974 controls (53.7)[7]	2,576 cases and 60,771 controls (4.1%)	0.68 (0.67–0.69)	0.017

Table S2: The AUC values (reported in Khera et al.[2]) and R^2 values converted from the AUC given the sample size, prevalence in discovery and testing datasets. R^2 values were converted from the AUC using the well-established theory[8; 9].

Supplemental References

1. Olkin, I., and Finn, J.D. (1995). Correlations redux. Psychological Bulletin 118, 155.
2. 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.
3. Nikpay, M., Goel, A., Won, H.-H., Hall, L.M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C.P., and Hopewell, J.C. (2015). A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. Nature Genetics 47, 1121.
4. Christophersen, I.E., Rienstra, M., Roselli, C., Yin, X., Geelhoed, B., Barnard, J., Lin, H., Arking, D.E., Smith, A.V., and Albert, C.M. (2017). Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. Nat Genet 49, 946-952.
5. Scott, R.A., Scott, L.J., Mägi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., and Eicher, J.D. (2017). An expanded genome-wide association study of type 2 diabetes in Europeans. Diabetes 66, 2888-2902.
6. Liu, J.Z., Van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., and Shah, T. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature genetics 47, 979-986.
7. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., and Rostamianfar, A. (2017). Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92-94.
8. 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.
9. 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.