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

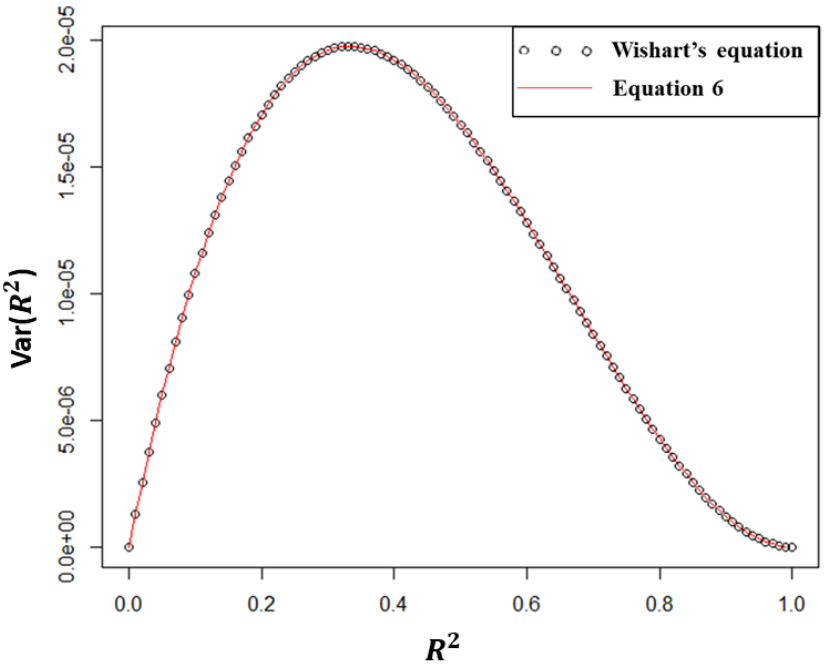

Figure S1: Wishart's equation and Equation 6 provide identical results for any R^2 values ranging from 0 to 1 using sample size 25000.

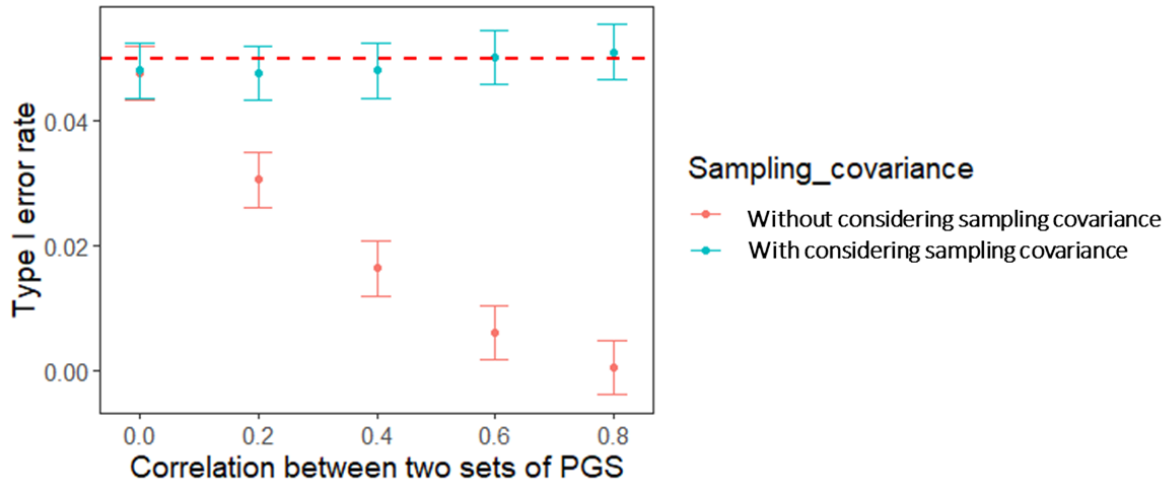

Figure S2: Incorrect type I error rate for testing the difference between two R^2 values (r^2_{y,x_1} vs. r^2_{y,x_2}) when ignoring the correlation between two sets of PGS (r_{x_1,x_2}> 0). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.100 & 0.100 \\ 0.100 & 1 & various \\ 0.100 & various & 1 \end{bmatrix}$, 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).

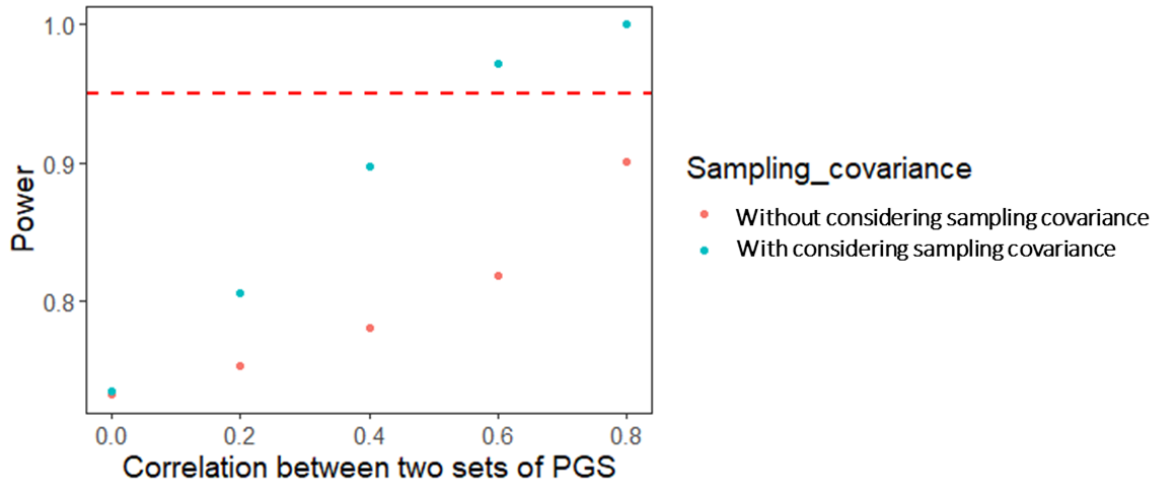

Figure S3: Reduced power for testing the difference between two R^2 values (r^2_{y,x_1} vs. r^2_{y,x_2}) when ignoring the correlation between two sets of PGS (r_{x_1,x_2} > 0). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.120 & 0.100 \\ 0.120 & 1 & various \\ 0.100 & various & 1 \end{bmatrix}$, 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.

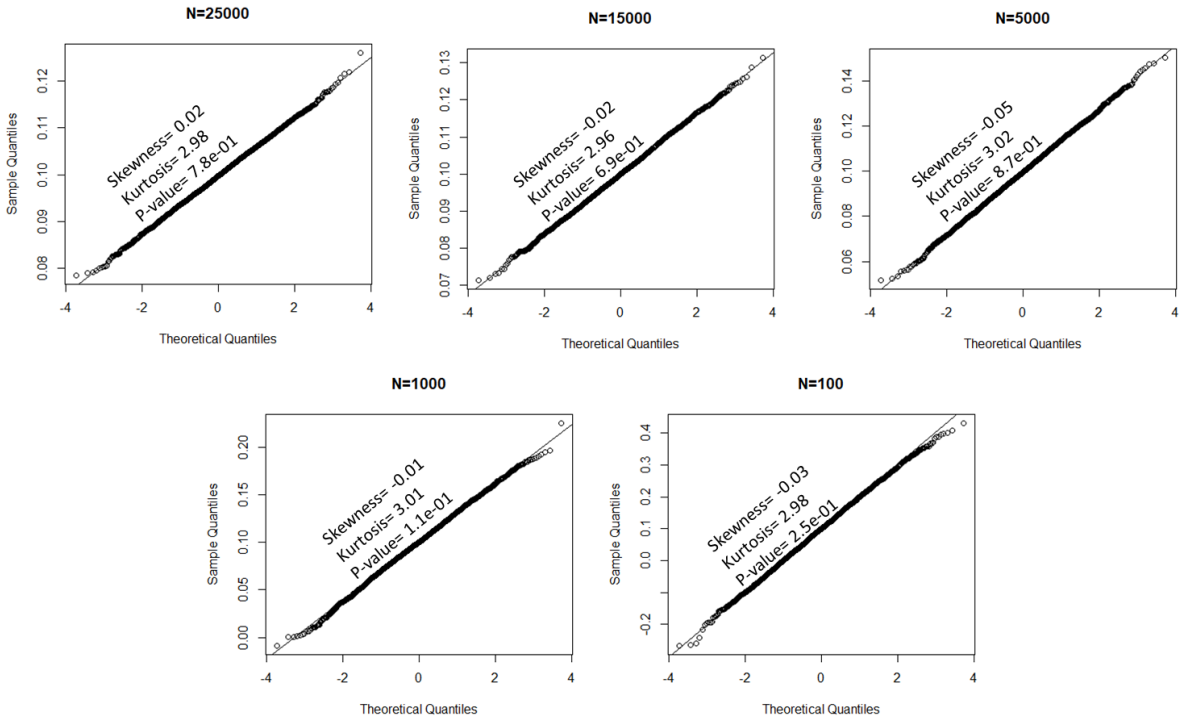

Figure S4: Distribution of regression coefficients are asymptotically normal when correlation between two PGS is 0.10. Simulations of y and x_1 were based on a correlation of $r_{y,x_1} = 0.10$ and the regression coefficient was estimated from a model $y = x_1 + e$ using 5,000 replications. The sample 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).

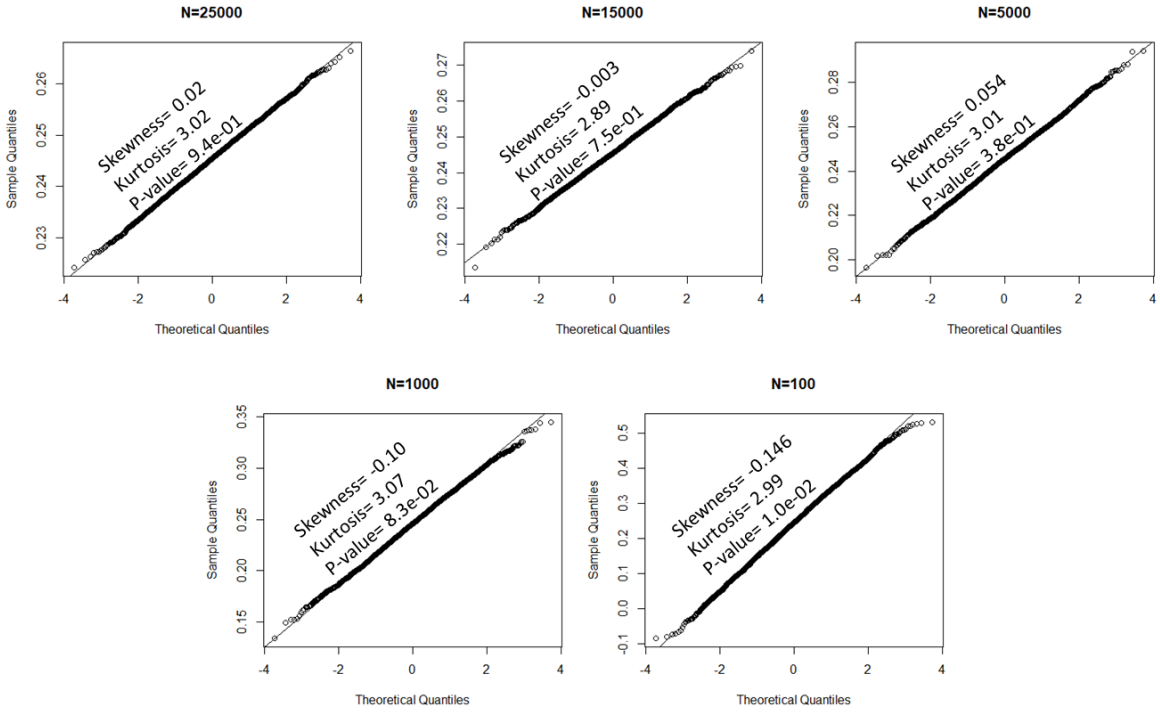

Figure S5: Distribution of regression coefficients are asymptotically normal when correlation between two PGS is 0.25. Simulations of y and x_1 were based on a correlation of $r_{y,x_1} = 0.25$ and the regression coefficient was estimated from a model $y = x_1 + e$ using 5,000 replications. The sample 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.

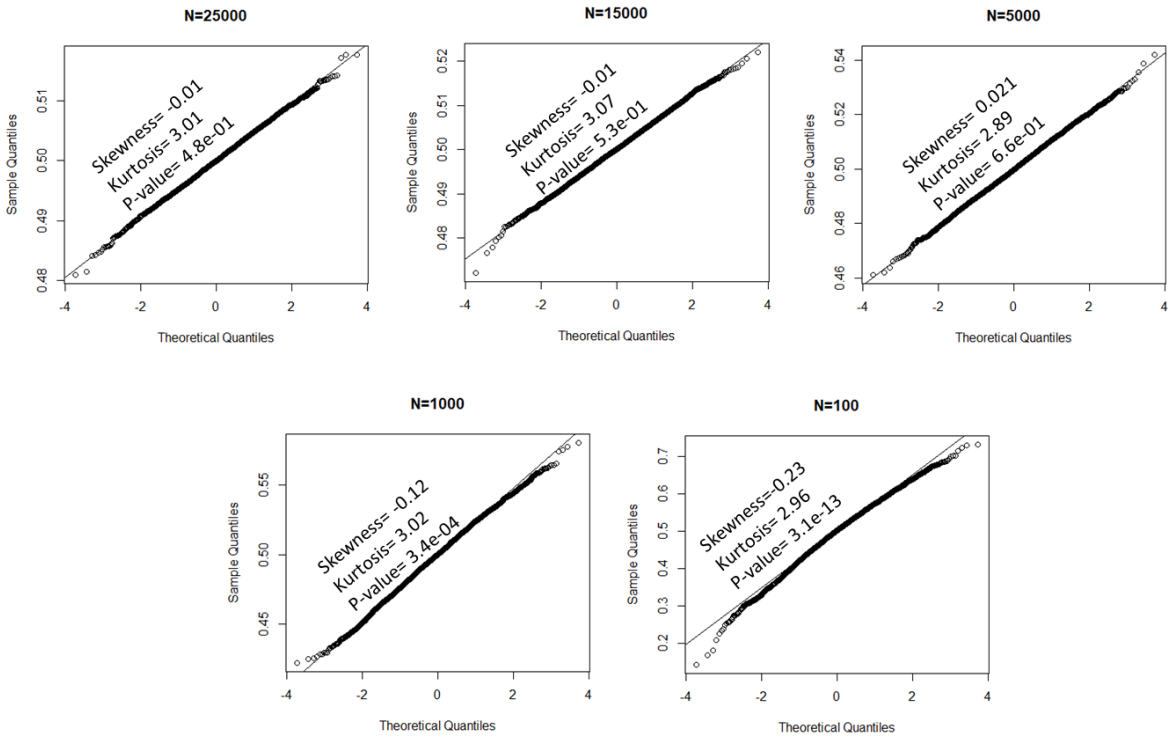

Figure S6: Distribution of regression coefficients are asymptotically normal when correlation between two PGS is 0.50. Simulations of y and x_1 were based on a correlation of $r_{y,x_1} = 0.50$ and the regression coefficient was estimated from a model $y = x_1 + e$ using 5,000 replications. The sample 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.

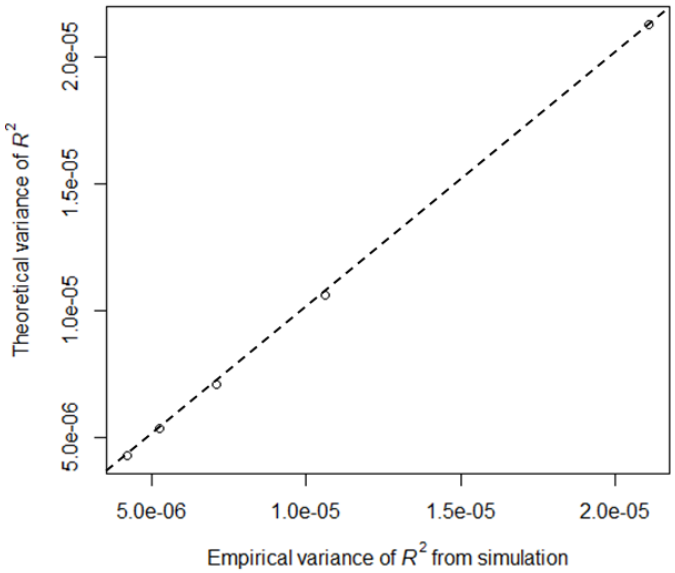

Figure S7: Nearly identical values between the theoretical and empirical variances of R^2 (r^2_{y,x_1}) estimated from 10,000 simulated replicates when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.246 & 0.139 \\ 0.246 & 1 & 0.315 \\ 0.139 & 0.315 & 1 \end{bmatrix}$ 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.

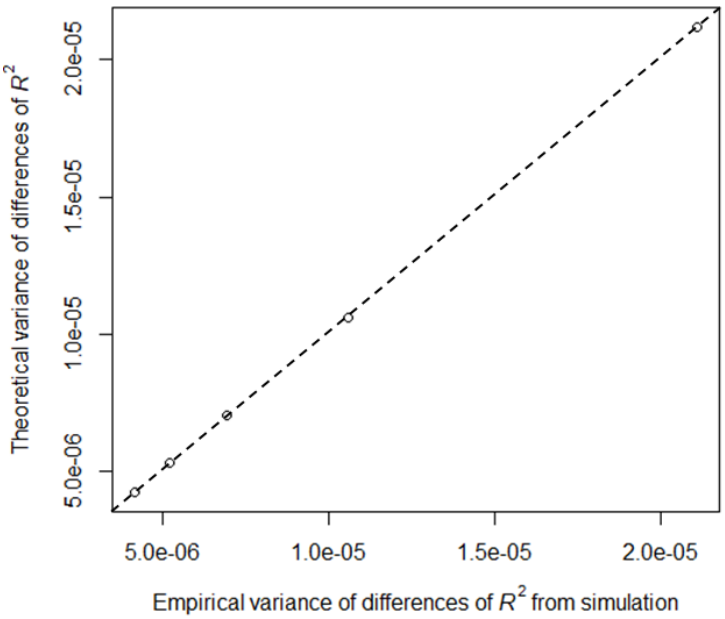

Figure S8: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,x_1} - r^2_{y,x_2}$) estimated from 10,000 simulated replicates when varying sample size.

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.

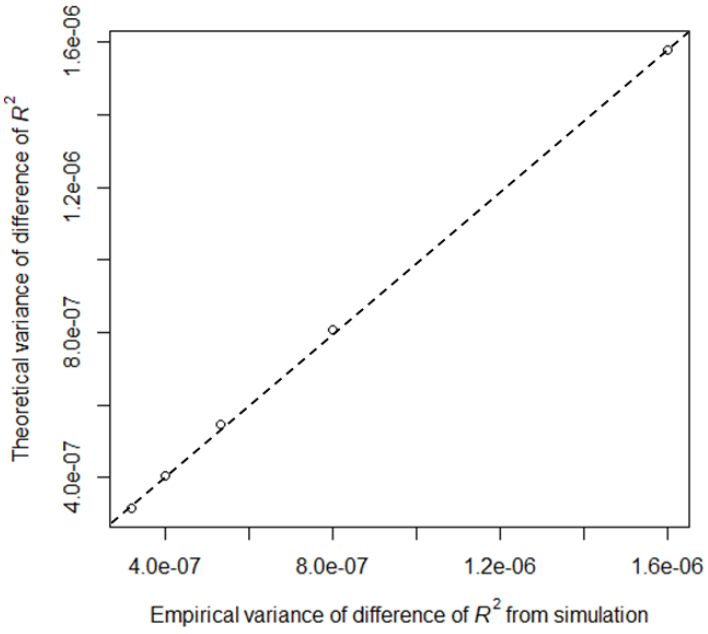

Figure S9: Nearly identical values between the theoretical and empirical variances of R^2 difference of ($r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$) estimated from 10,000 simulated replicates when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.246 & 0.139 \\ 0.246 & 1 & 0.315 \\ 0.139 & 0.315 & 1 \end{bmatrix}$, 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.

.

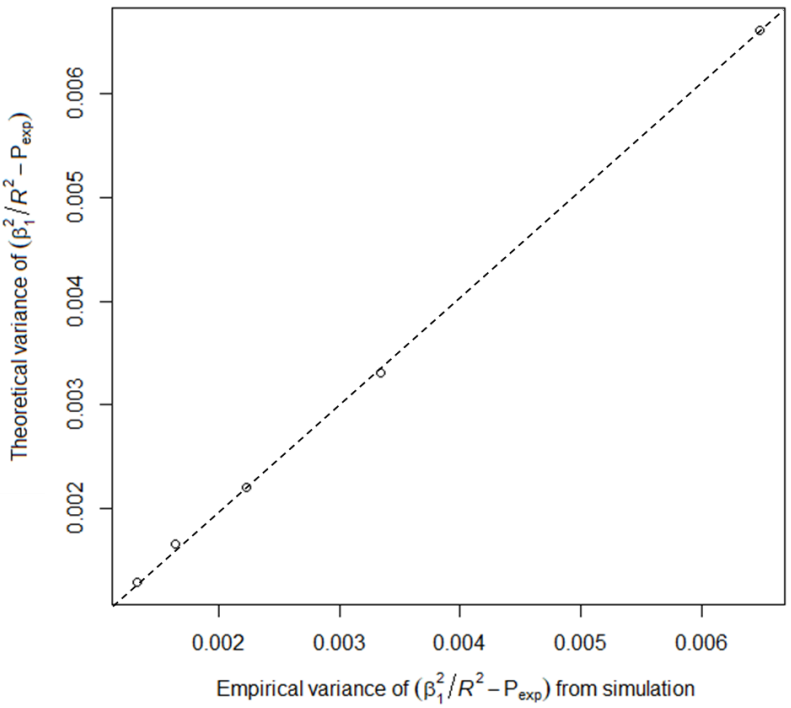

Figure S10: Nearly identical values between the theoretical and empirical variances of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ estimated from 10,000 simulated replicates when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.176 & 0.148 \\ 0.176 & 1 & 0.610 \\ 0.148 & 0.610 & 1 \end{bmatrix}$, 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.

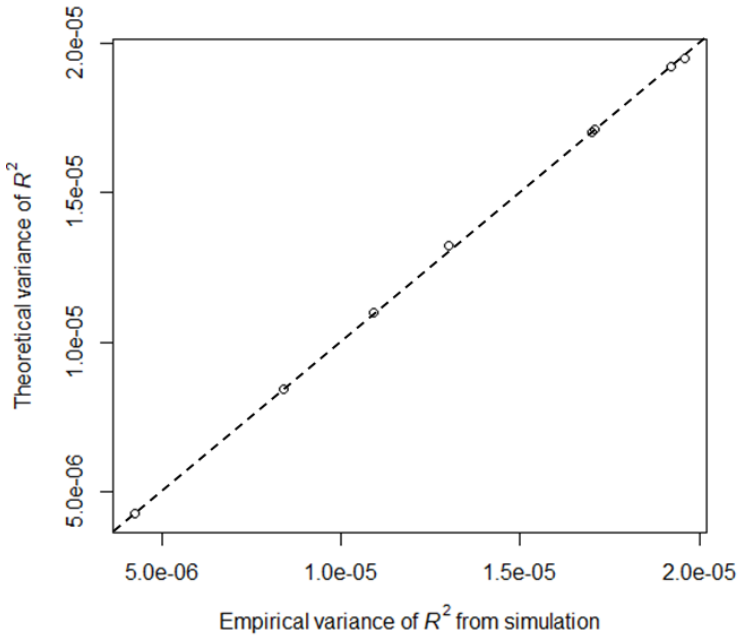

Figure S11: Nearly identical values between the theoretical and empirical variances of R^2 (r^2_{y,x_1}) estimated from 10,000 simulated replicates when varying R^2 value. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.447 \\ various & 1 & 0.800 \\ 0.447 & 0.800 & 1 \end{bmatrix}$ 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.

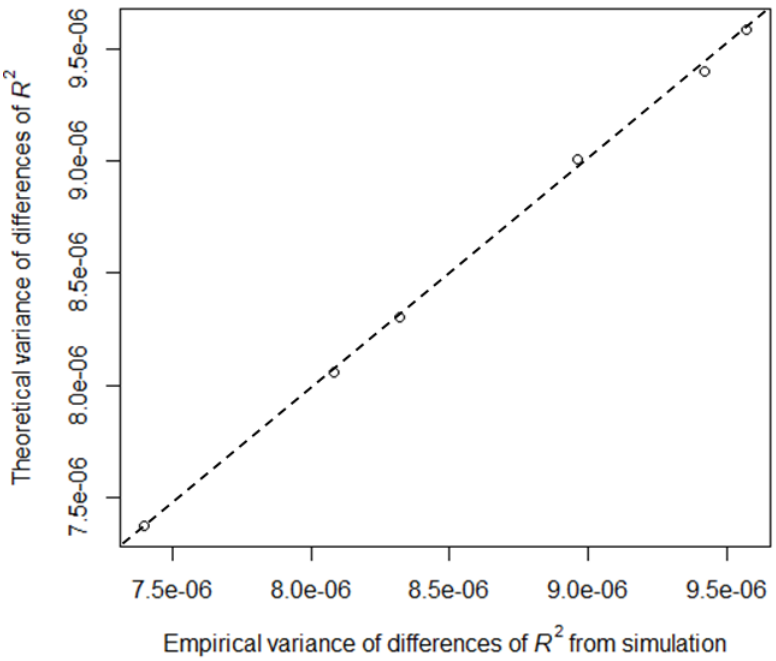

Figure S12: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,x_1} - r^2_{y,x_2}$) estimated from 10,000 simulated replicates when varying R^2 difference.

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.

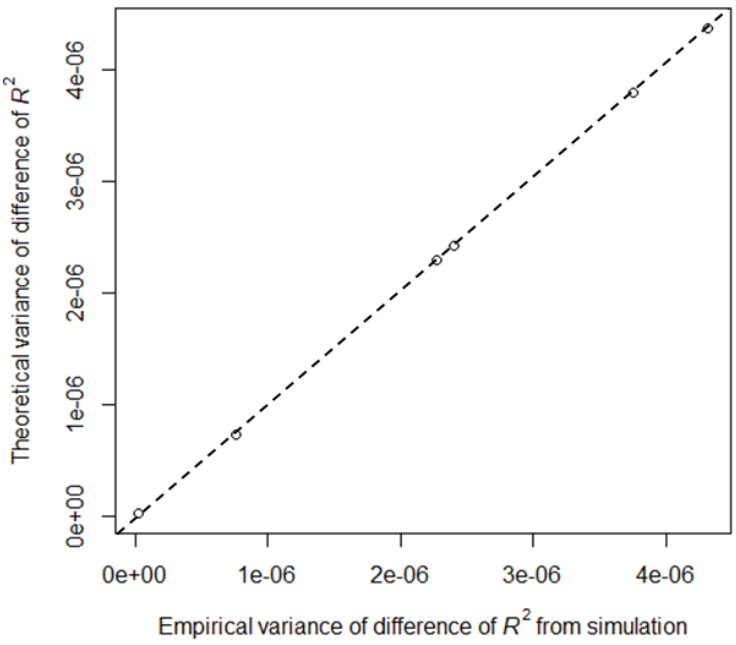

Figure S13: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$) estimated from 10,000 simulated replicates when varying R^2 difference. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.447 \\ various & 1 & 0.800 \\ 0.447 & 0.800 & 1 \end{bmatrix}$, 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.

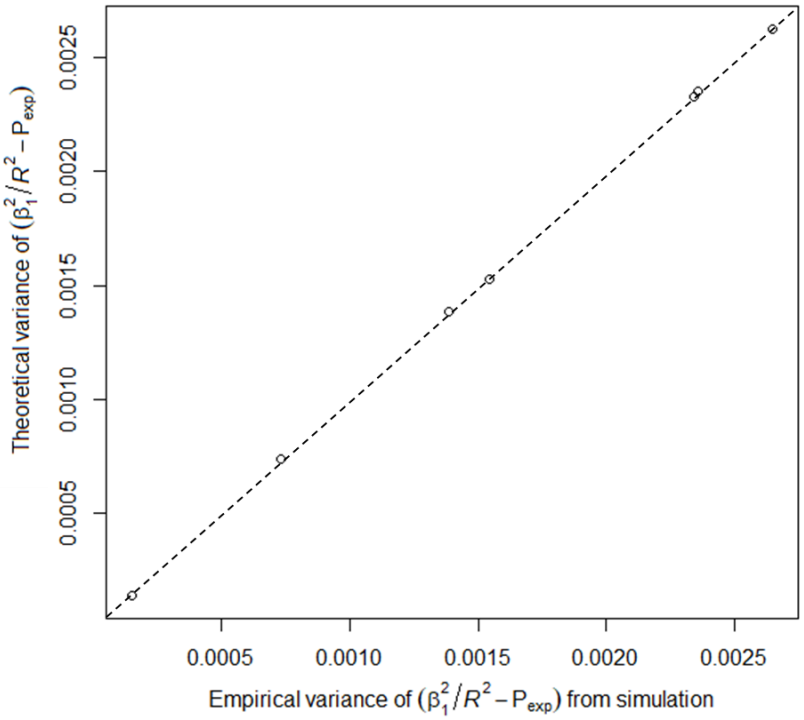

Figure S14. Nearly identical values between the theoretical and empirical variances of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ estimated from 10,000 simulated replicates when varying correlation structure. Simulations of *y*, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.148 \\ various & 1 & 0.610 \\ 0.148 & 0.610 & 1 \end{bmatrix}$, 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).

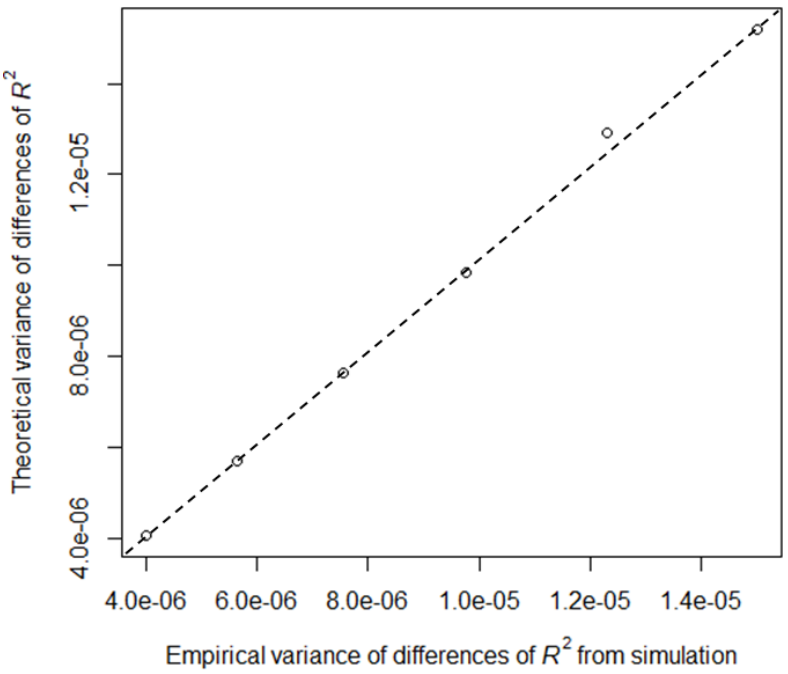

Figure S15: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y_1,x_1} - r^2_{y_2,x_2}$) estimated from 10,000 simulated replicates when using two sets of independent PGS (e.g., Male vs female PGS). Simulations of y_1 and x_1 were based on a correlation 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.

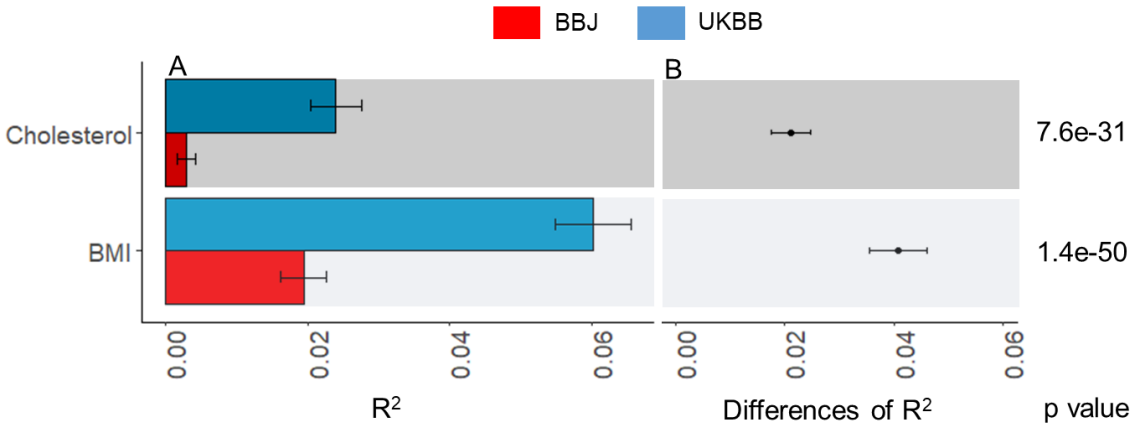

Figure S16: The predictive ability of (R^2) for male and female, when predicting European male and female separately using UKBB and BBJ discovery samples.
Panel A: The main bars represent R^2 values and error bars correspond 95% confidence intervals.
Panel B: Dot points represent the differences of R^2 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).

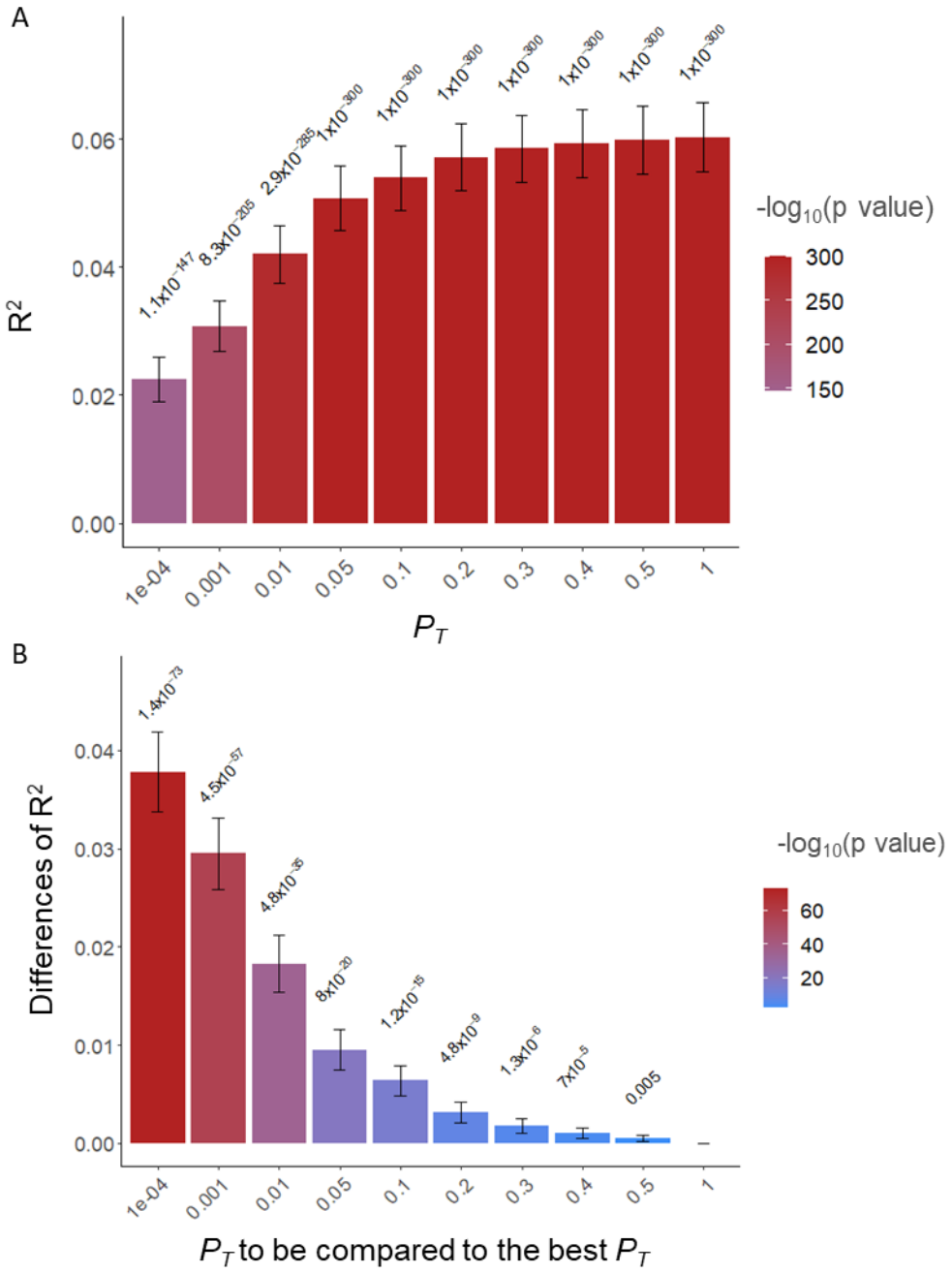

Figure S17: The predictive ability (R^2) of PGS estimated based on SNPs below the p_T when predicting BMI in 28,880 European samples using UKBB discovery samples (GWAS summary statistics).

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.

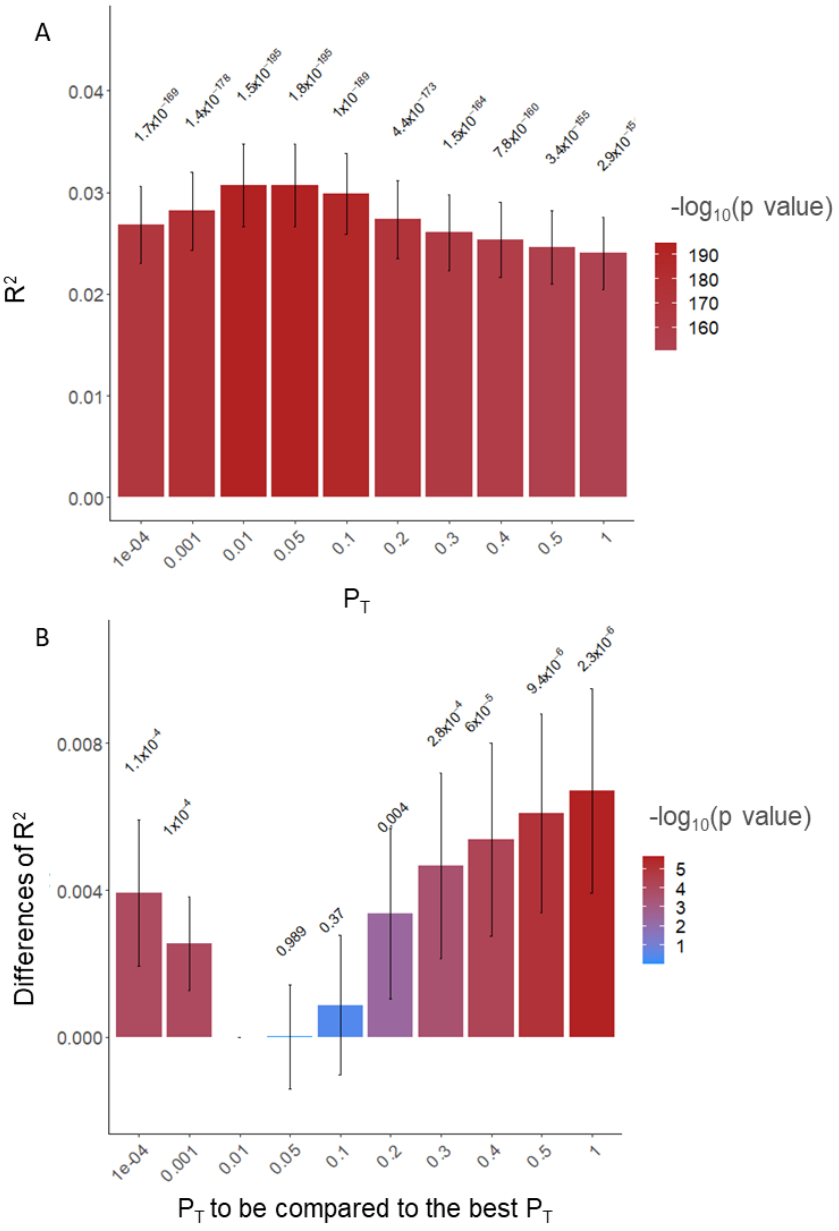

Figure S18: The predictive ability (R^2) of PGS estimated based on SNPs below the p_T when predicting cholesterol in 28,880 European samples using UKBB discovery samples (GWAS summary statistics).

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.

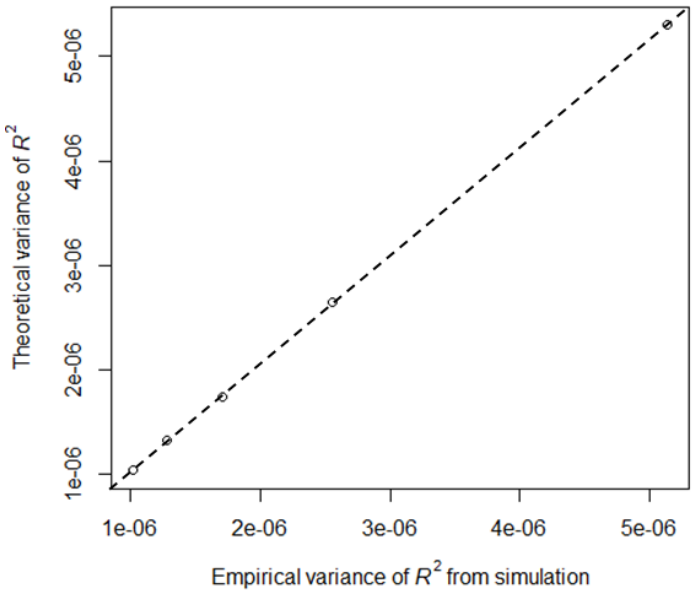

Figure S19: Nearly identical values between the theoretical and empirical variances of R^2 (r^2_{y,x_1}) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.246 & 0.139 \\ 0.246 & 1 & 0.315 \\ 0.139 & 0.315 & 1 \end{bmatrix}$ 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.

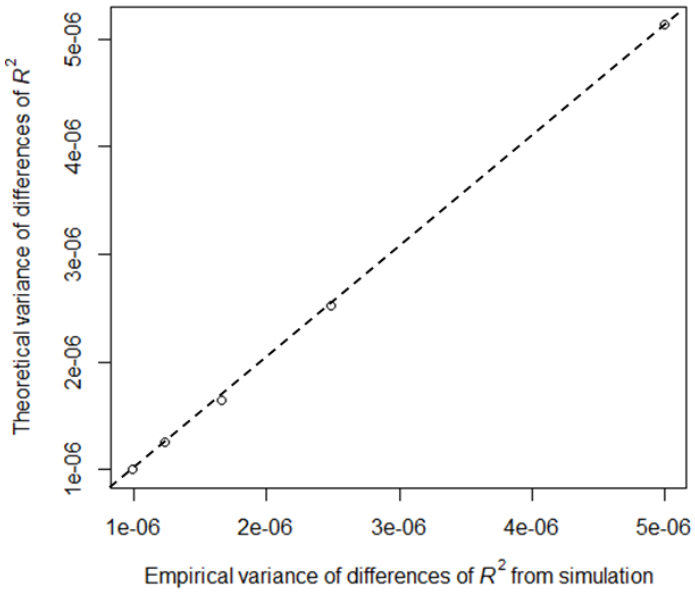

Figure S20: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,x_1} - r^2_{y,x_2}$) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation 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.

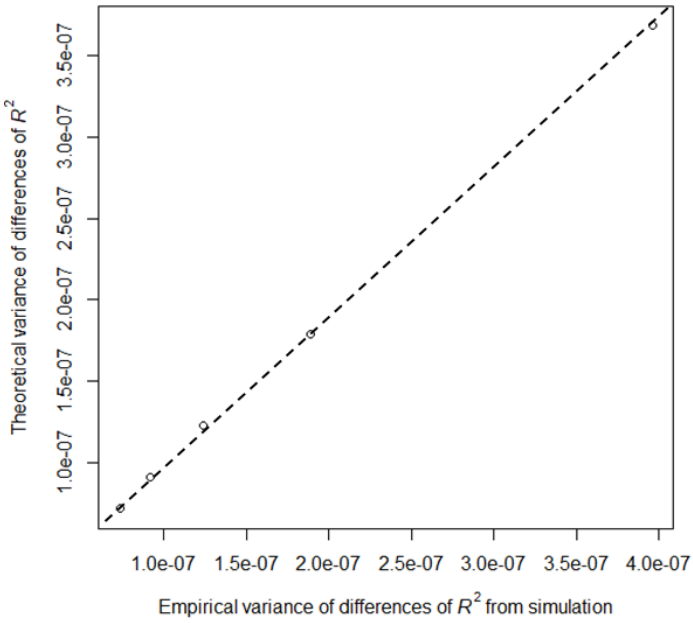

Figure S21: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying sample size. Simulations of y, x_1 and x_2 were based on 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.

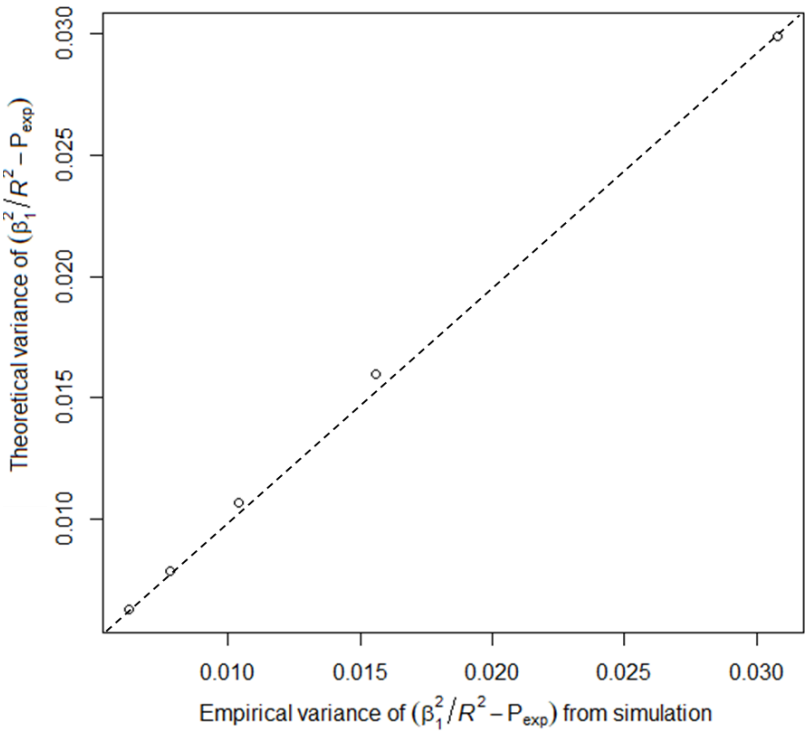

Figure 22: Nearly identical values between the theoretical and empirical variances of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying sample size. Simulations of y, x_1 and x_2 were based on a correlation 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.

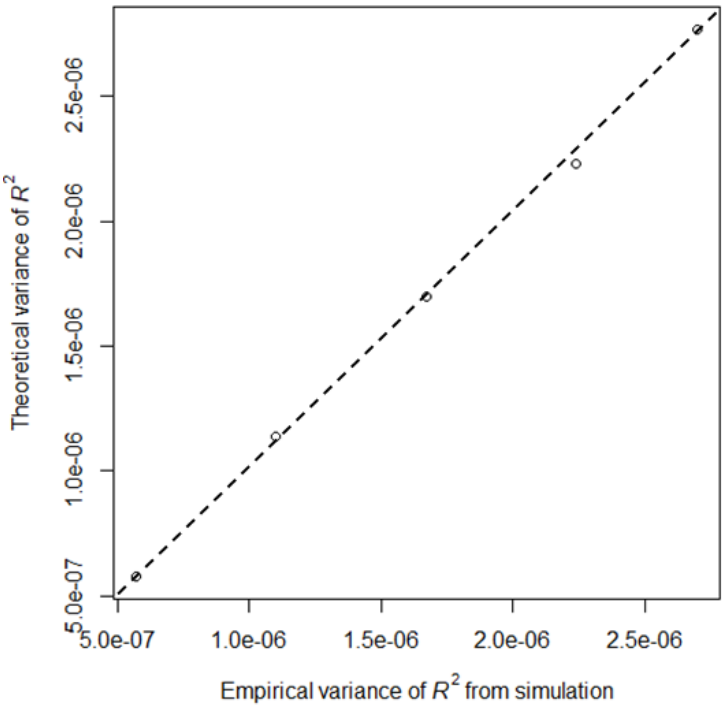

Figure S23: Nearly identical values between the theoretical and empirical variances of R^2 (r^2_{y,x_1}) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying R^2 value. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.141 \\ various & 1 & 0.800 \\ 0.141 & 0.8 & 1 \end{bmatrix}$ 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.

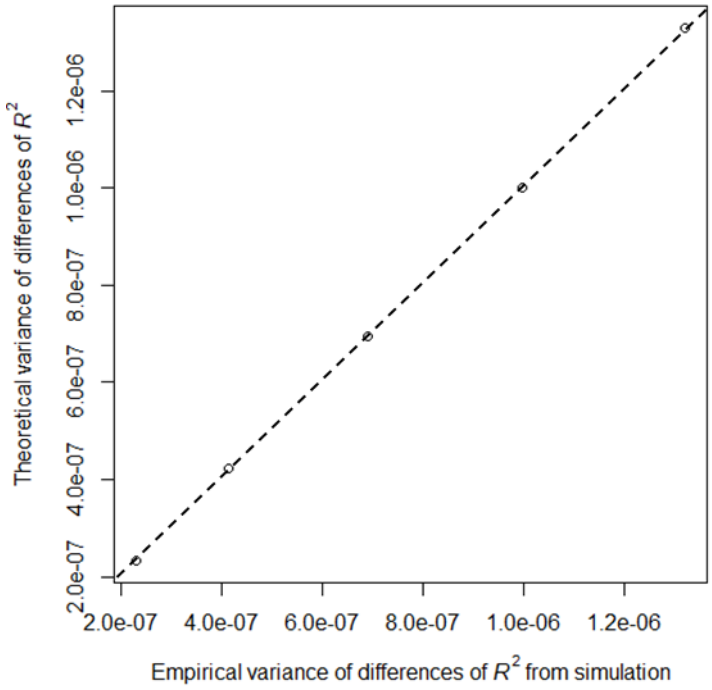

Figure S24: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,x_1} - r^2_{y,x_2}$) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying R^2 difference. Simulations of *y*, x_1 and x_2 were based on 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.

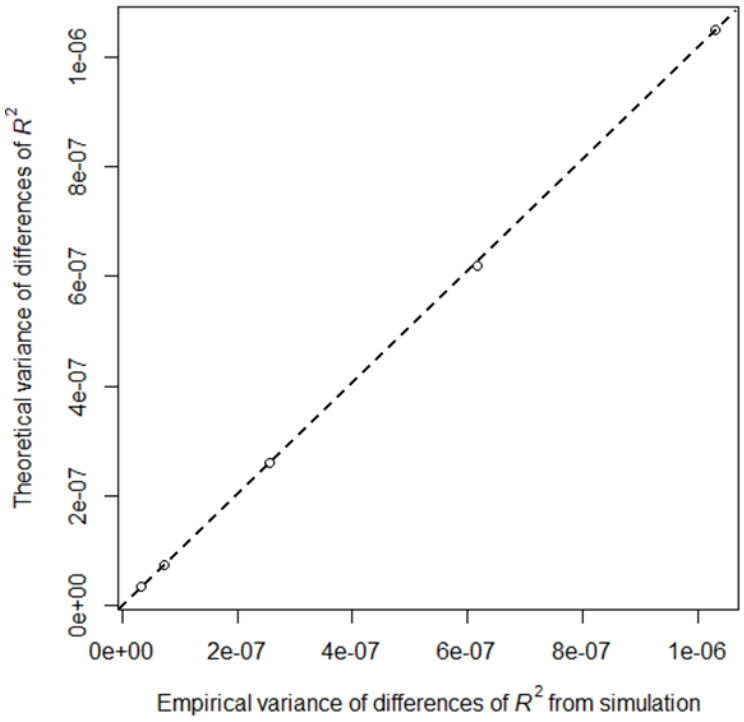

Figure S25: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$) estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying R^2 difference. Simulations of y, x_1 and x_2 were based 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.

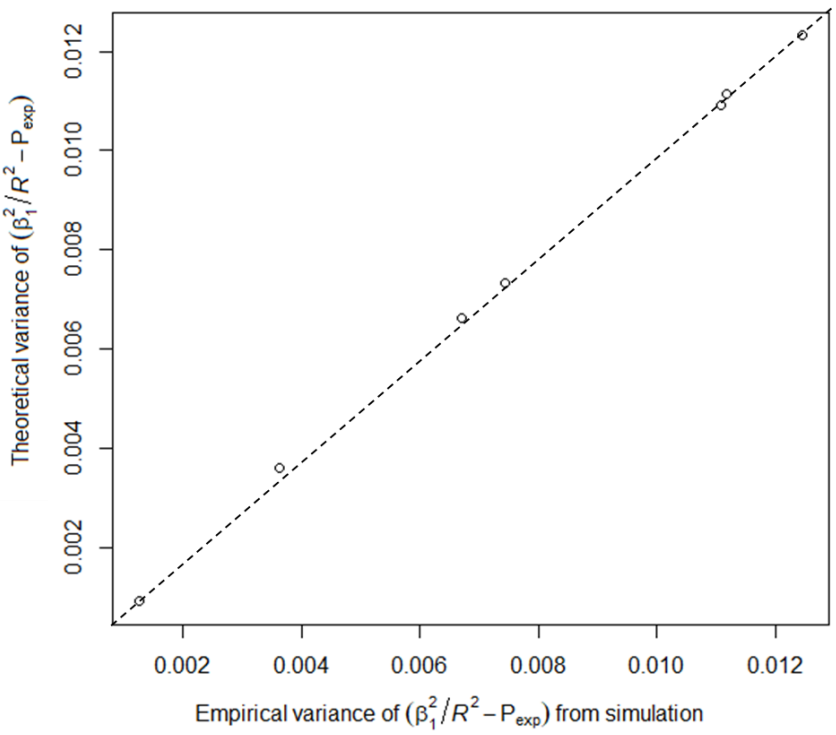

Figure S26: Nearly identical values between the theoretical and empirical variances of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ estimated from 10,000 simulated replicates of binary responses assuming 5% disease prevalence when varying correlation structure. Simulations of *y*, x_1 and x_2 were based on 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).

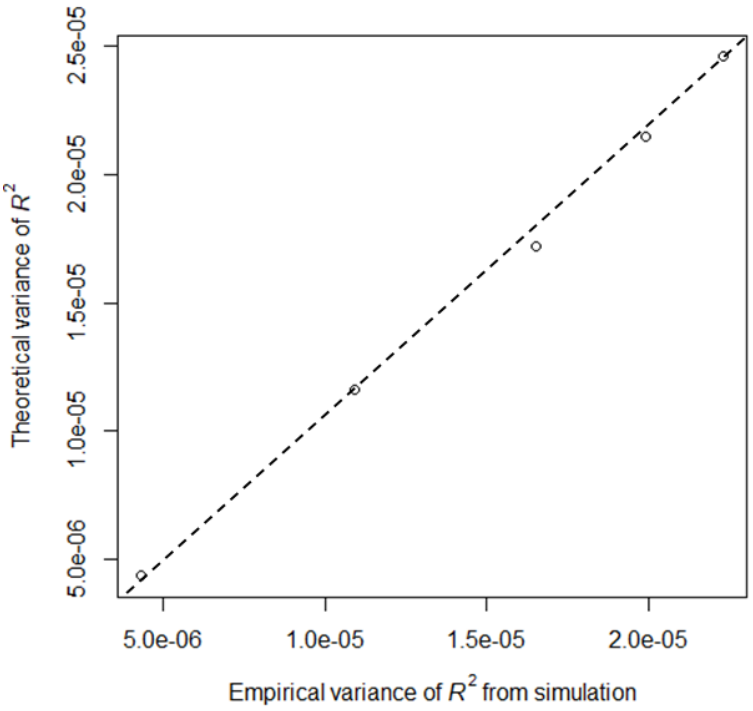

Figure S27: Nearly identical values between the theoretical and empirical variances of R^2 (r^2_{y,x_1}) estimated from 10,000 simulated replicates of ascertained case-control (10000 cases and 10000 controls) assuming 5% disease prevalence and 20000 individuals. Simulations of y, x_1 and x_2 were based 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.

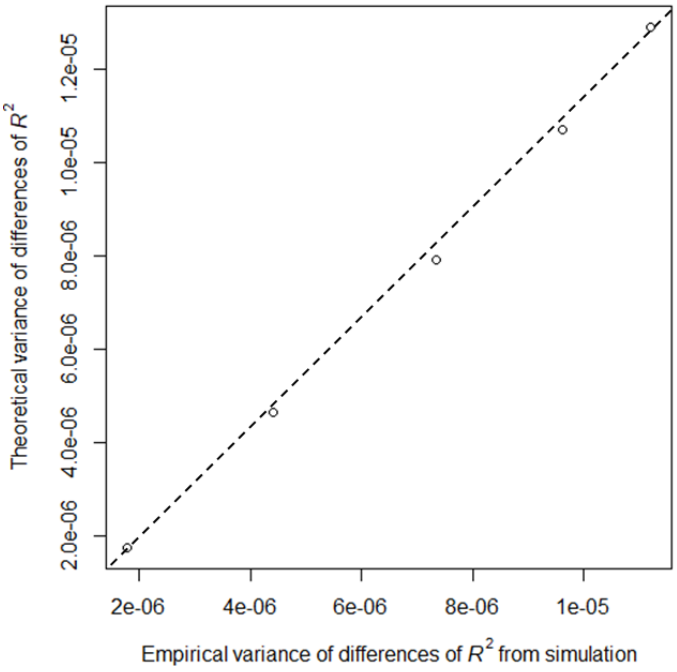

Figure S28: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,x_1} - r^2_{y,x_2}$) estimated from 10,000 simulated replicates of ascertained case-control (10000 cases and 10000 controls) assuming 5% disease prevalence and 20000 individuals. Simulations 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.

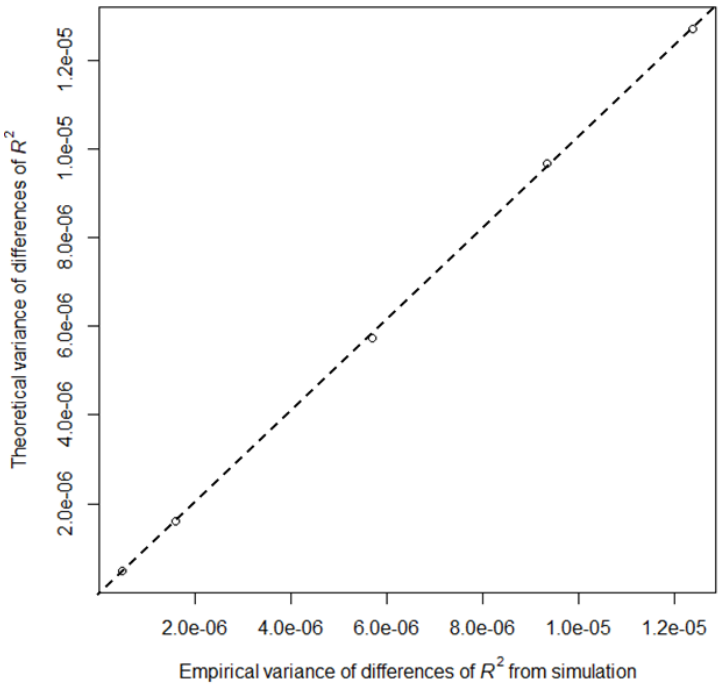

Figure S29: Nearly identical values between the theoretical and empirical variances of R^2 difference ($r^2_{y,(x_1,x_2)} - r^2_{y,x_1}$) estimated from 10,000 simulated replicates of ascertained case-control (10000 cases and 10000 controls) assuming 5% disease prevalence and 20000 individuals.

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.

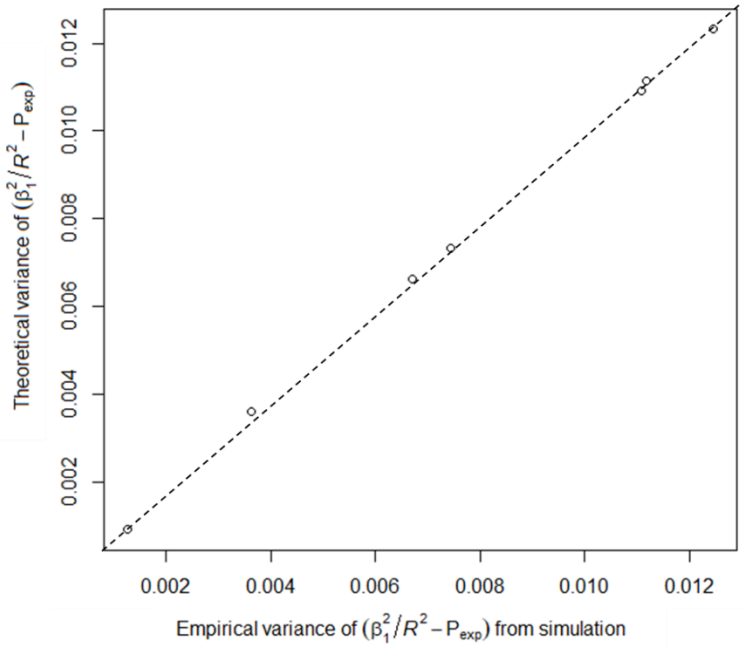

Figure S30: Nearly identical values between the theoretical and empirical variances of $\frac{\hat{\beta}_1^2}{R^2} - P_{exp}$ estimated from 10,000 simulated replicates of ascertained case-control (10000 cases and 10000 controls) assuming 5% disease prevalence and 20000 individuals. Simulations of y, 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).

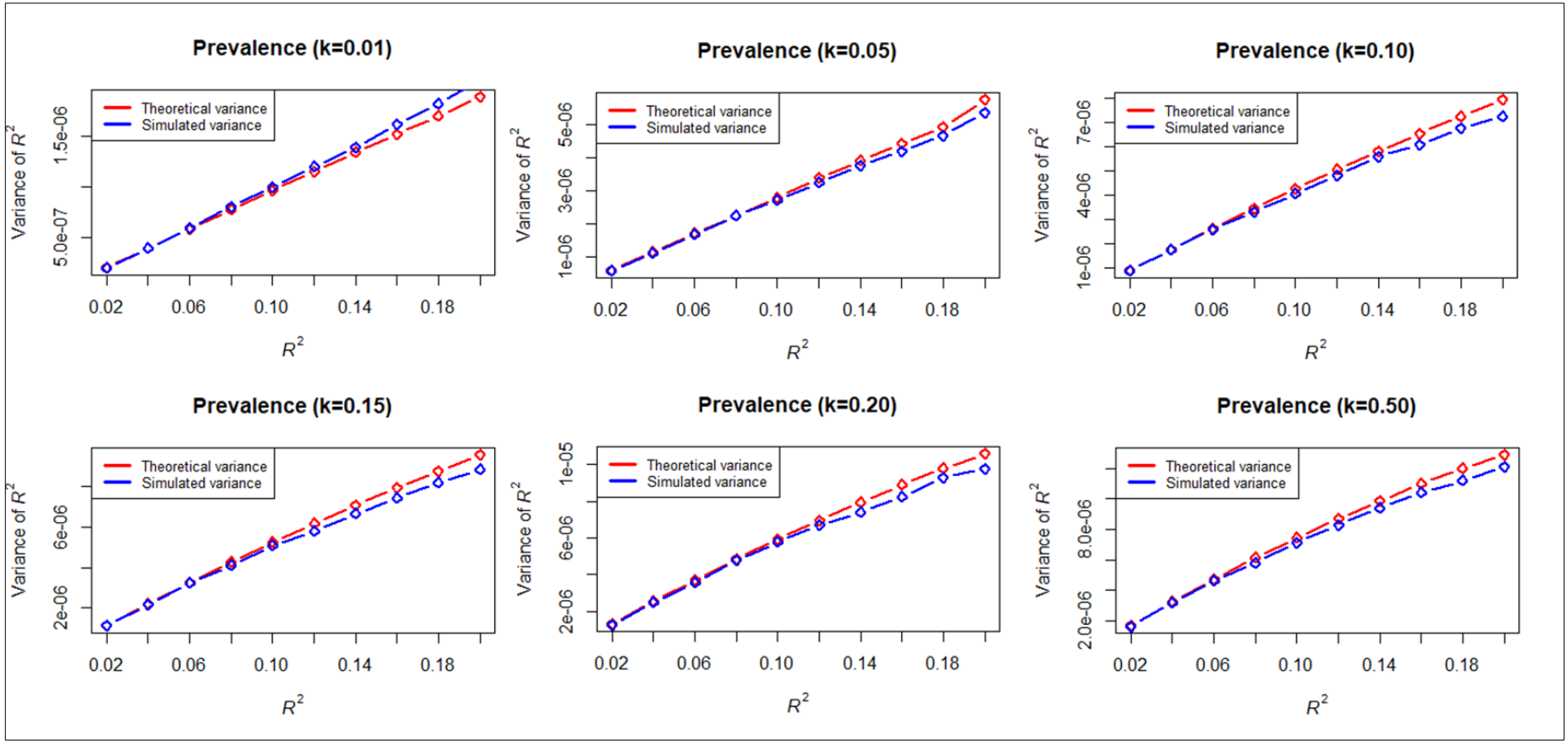

Figure S31: The empirical and theoretical variances diverge when R^2 values are more than 0.1 for binary responses, noting that $R^2 > 0.1$ is not frequently observed (see Supplemental Table 2). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & various & 0.141 \\ various & 1 & 0.800 \\ 0.141 & 0.8 & 1 \end{bmatrix}$ 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.

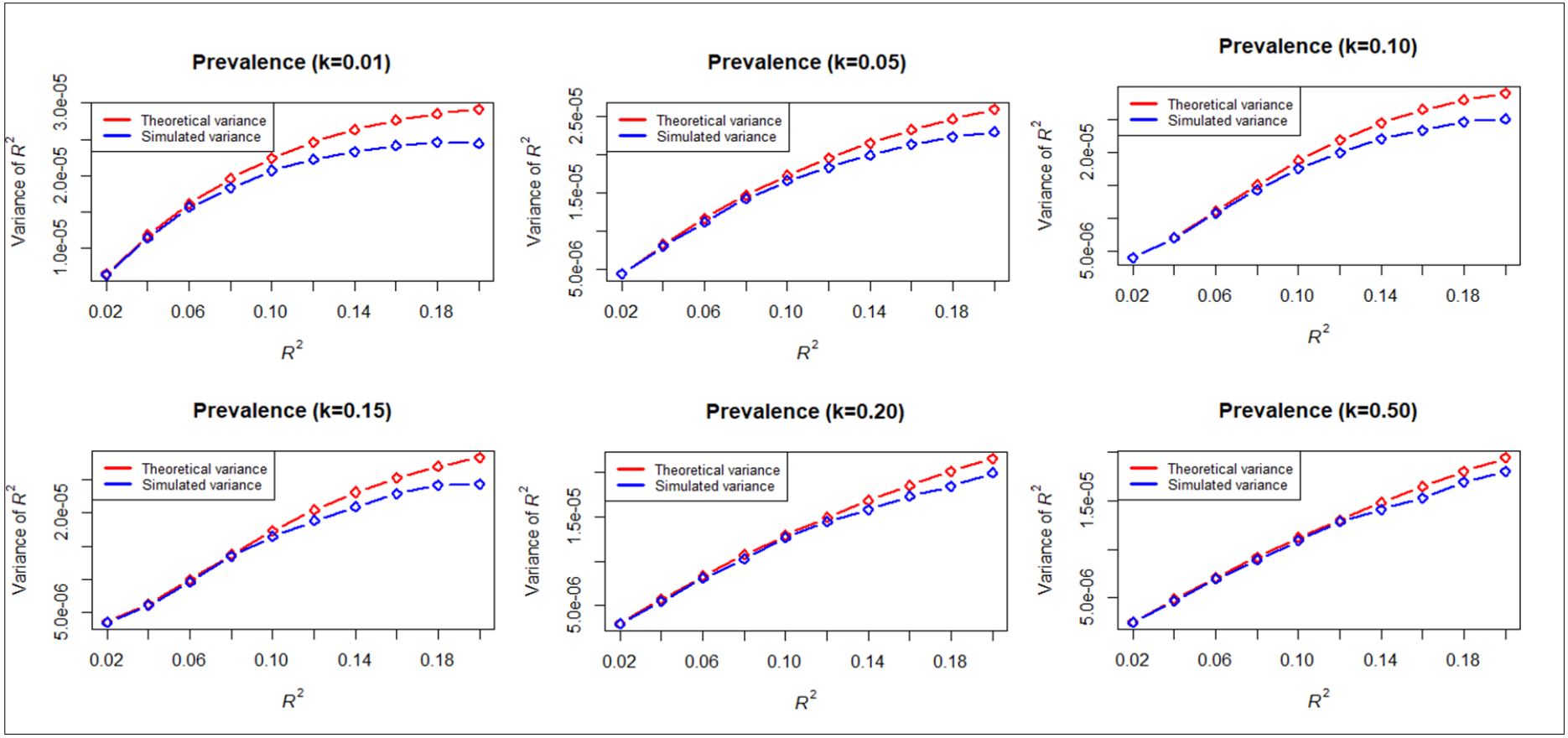

Figure S32: The empirical and theoretical variances become disagreed when R^2 values are more than 0.1 for ascertained case-control samples in the reference dataset (10000 cases and 10000 controls), noting that R^2 > 0.1 is not frequently observed (see Supplemental Table 2). Simulations of y, x_1 and 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.

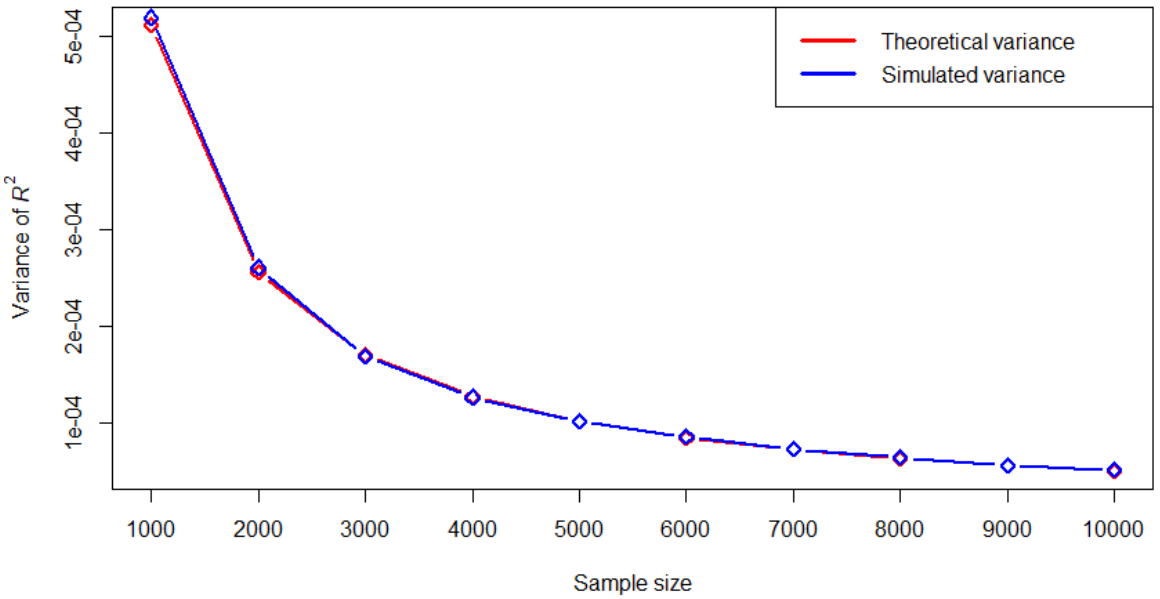

Figure S33: The empirical and theoretical variances agree even with sample size 2000 for quantitative phenotypes. Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.44 & 0.31 \\ 0.44 & 1 & 0.800 \\ 0.31 & 0.8 & 1 \end{bmatrix}$ 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.

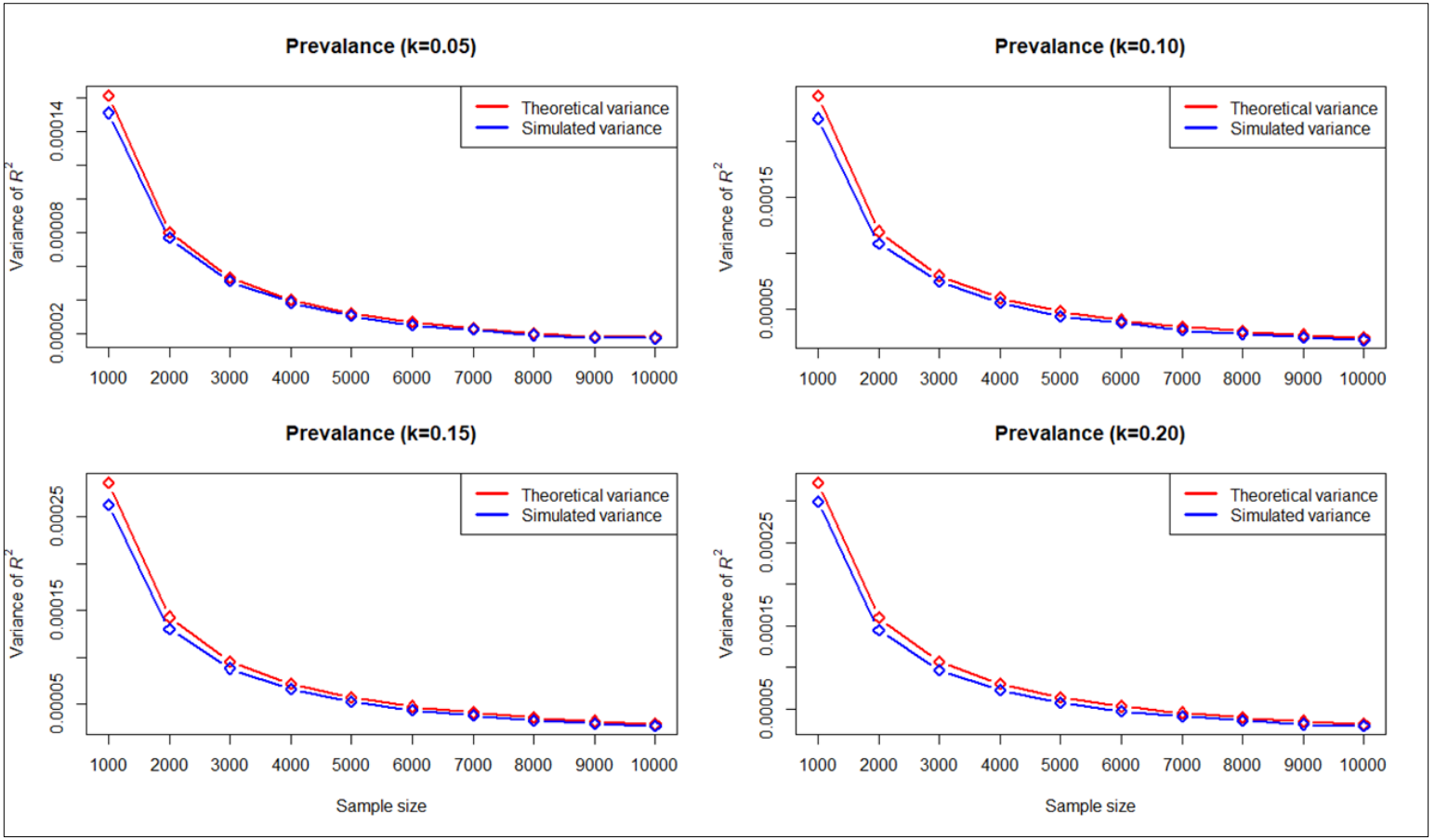

Figure S34: The empirical and theoretical variances become disagreed when sample size is < 5000 for binary responses under scenario of different prevalence rate (k). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.44 & 0.31 \\ 0.44 & 1 & 0.800 \\ 0.31 & 0.8 & 1 \end{bmatrix}$ 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.

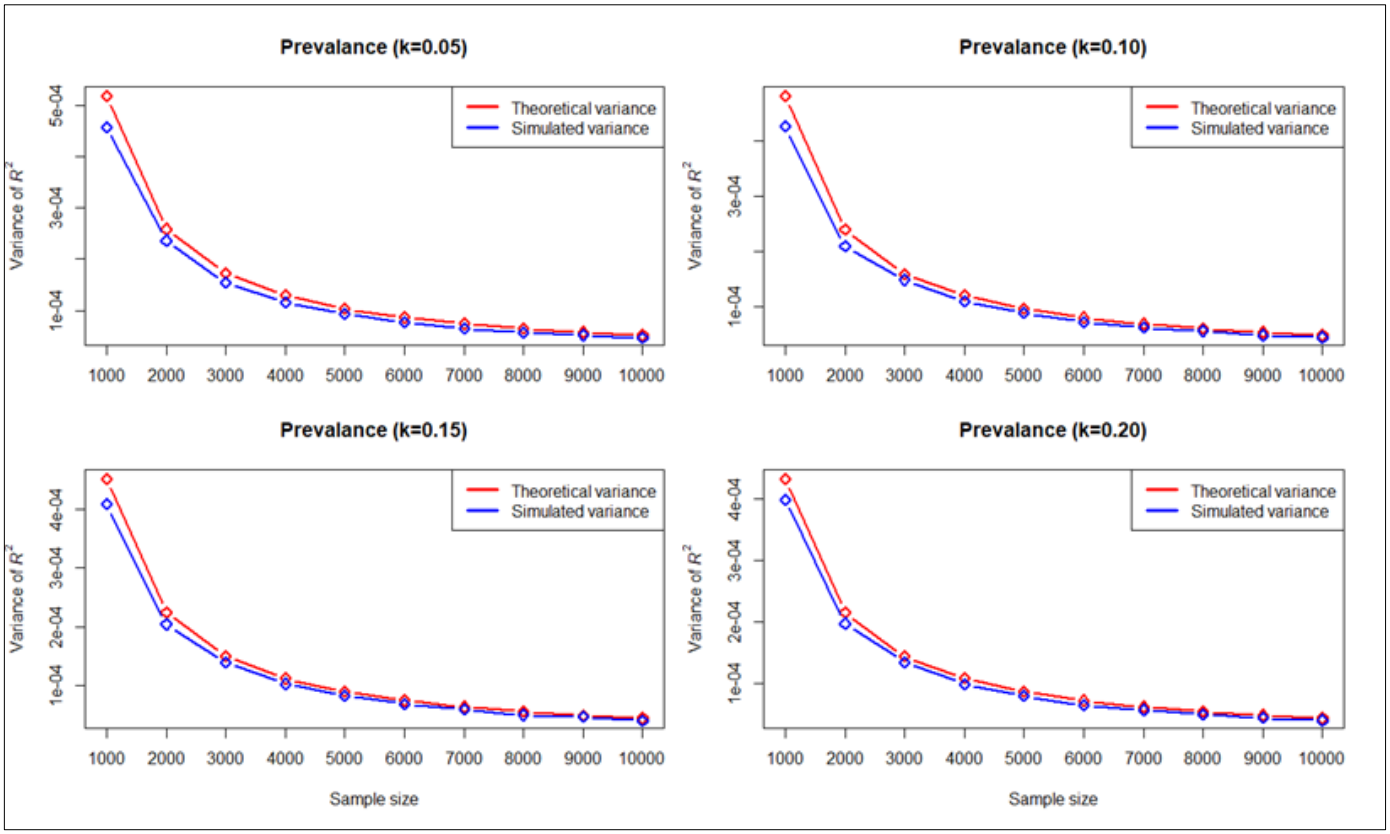

Figure S35: The empirical and theoretical variances become disagreed when sample size is < 5000 for ascertained case-control samples in the reference dataset (50% cases and 50% controls) under scenario of different prevalence rate (k). Simulations of y, x_1 and x_2 were based on a correlation structure $\begin{bmatrix} 1 & r_{y,x_1} & r_{y,x_2} \\ r_{y,x_1} & 1 & r_{x_1,x_2} \\ r_{y,x_2} & r_{x_1,x_2} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.44 & 0.31 \\ 0.44 & 1 & 0.800 \\ 0.31 & 0.8 & 1 \end{bmatrix}$ 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.