Title

SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics.

Permalink

https://escholarship.org/uc/item/1v1714b9

Journal

Bioinformatics, 35(20)

ISSN

1367-4803

Authors

Song, Lei
Liu, Aiyi
Shi, Jianxin
et al.

Publication Date

2019-10-15

DOI

10.1093/bioinformatics/btz176

Peer reviewed

Bioinformatics, 35(20), 2019, 4038–4044
doi: 10.1093/bioinformatics/btz176
Advance Access Publication Date: 26 March 2019
Original Paper

Genetics and population analysis

SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics

Lei Song[1,2], Aiyi Liu[3], Molecular Genetics of Schizophrenia Consortium† and Jianxin Shi[1,]*

[1]Biostatistics Branch, [2]Cancer Genomics Research Laboratory, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA and [3]Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research, National Institute of Child Health and Human Development, Bethesda, MD 20817, USA

*To whom correspondence should be addressed.
†Details of the MGS Consortium are listed in Acknowledgements.

Associate Editor: Russell Schwartz

Received on June 6, 2018; revised on February 8, 2019; editorial decision on March 8, 2019; accepted on March 22, 2019

Abstract

Motivation: Polygenic risk score (PRS) methods based on genome-wide association studies (GWAS) have a potential for predicting the risk of developing complex diseases and are expected to become more accurate with larger training datasets and innovative statistical methods. The area under the ROC curve (AUC) is often used to evaluate the performance of PRSs, which requires individual genotypic and phenotypic data in an independent GWAS validation dataset. We are motivated to develop methods for approximating AUC of PRSs based on the summary level data of the validation dataset, which will greatly facilitate the development of PRS models for complex diseases.

Results: We develop statistical methods and an R package SummaryAUC for approximating the AUC and its variance of a PRS when only the summary level data of the validation dataset are available. SummaryAUC can be applied to PRSs with SNPs either genotyped or imputed in the validation dataset. We examined the performance of SummaryAUC using a large-scale GWAS of schizophrenia. SummaryAUC provides accurate approximations to AUCs and their variances. The bias of AUC is typically <0.5% in most analyses. SummaryAUC cannot be applied to PRSs that use all SNPs in the genome because it is computationally prohibitive.

Availability and implementation: https://github.com/lsncibb/SummaryAUC.

Contact: Jianxin.Shi@nih.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Large-scale genome-wide association studies (GWAS) have identified dozens or even hundreds of common SNPs associated with many complex diseases, including psychiatric conditions, e.g. schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), type 2 diabetes (Scott *et al.*, 2017) and common cancers, e.g. breast cancer (Michailidou *et al.*, 2017) and prostate cancers (Al Olama *et al.*, 2014). Heritability analysis using algorithms such as genome-wide complex trait analysis (GCTA) (Yang *et al.*, 2010) and LD-score regressions (Bulik-Sullivan *et al.*, 2015) have shown that, for many complex diseases, common SNPs have the potential to explain substantially larger fraction of the phenotypic variance than that based on the established GWAS SNPs, suggesting a great promise for genetic risk prediction. In fact, polygenic risk scores (PRSs) have been proven useful for predicting complex disease risk and are expected to become more accurate with large training datasets (Chatterjee *et al.*, 2013; Dudbridge, 2013) and innovative statistical methods. For example, a PRS for schizophrenia based on tens of thousands of SNPs achieves an impressive prediction accuracy with an area under the receiver operating characteristic curve (AUC) of 0.75 (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) with ∼35 000 cases and 45 000 controls.

Much effort has been invested on developing PRS prediction models and on investigating the factors that determine the performance of PRSs. When raw genotypic and phenotypic data are available for the training dataset, machine learning algorithms (Kooperberg *et al.*, 2010; Wei *et al.*, 2009) and linear mixed models (Golan and Rosset, 2014; Maier *et al.*, 2015; Speed and Balding, 2014) can be used to develop PRSs. PRSs can also be constructed when only GWAS summary level data are available for training dataset by simple *P*-value thresholding (Purcell *et al.*, 2009) or by more sophisticated methods that model linkage disequilibrium (LD) (Vilhjalmsson *et al.*, 2015). We have recently extended the *P*-value thresholding method to further improve accuracy by accounting for winner's curse and by incorporating functional annotation data (Shi *et al.*, 2016). Different aspects of building PRS are discussed in a recent review paper (Chatterjee *et al.*, 2016).

Receiver operating characteristic (ROC) curve is one of the most popular tools for characterizing and comparing the diagnostic accuracy of binary classifier systems such as the PRSs in the present context. Since its first appearance in the Second World War for detecting enemy objects in battlefields, the ROC curve analysis has found its place in many others. A few books provide comprehensive coverage of the topics, see, among others (Hanley and Mcneil, 1982; Krzanowski and Hand, 2009; Pepe, 2003; Zou, 2011). The ROC curve of a PRS is generated by plotting its true positive rate against its false positive rate at various thresholds. The area under the ROC curve (AUC) provides a quantitative measure for the discrimination ability of a PRS (Hanley and Mcneil, 1982). AUC is the most frequently used quantitative measure for evaluating the discrimination performance of a PRS, although some concerns have been raised for using AUC as a criterion for model comparison and risk stratification (Katki and Schiffman, 2018). While AUC is defined as the area under the ROC curve, i.e. the integral of the curve, a more convenient mathematical expression of AUC is the probability that a randomly selected case has a larger PRS value than a randomly selected control. With this expression, one can show that an AUC estimator is closely related with the Mann–Whitney U statistic and the Wilcoxon rank test. Given PRS values for a set of cases and controls, one can easily calculate AUC using this approach and estimate the variance of the estimated AUC using bootstrap.

Calculating AUC for a PRS typically requires the individual level genotypic and phenotypic data in an independent GWAS validation dataset. One solution is to genotype a large set of subjects as a new validation GWAS dataset, which is financially expensive and time consuming. Another possibility is to request individual level data from existing large-scale GWAS independent of the training GWAS, which is also time consuming and may turn out to be infeasible because of data sharing policies. Instead, requesting summary statistics [odds ratio (OR), *P*-value and imputation quality for individual SNPs] from an independent validation GWAS consortium is much easier because such summary statistics are usually available online with open access. Thus, developing methods for evaluating the performance of PRSs based on the summary statistics of validation GWAS would substantially accelerate the assessment of PRSs for specific diseases and facilitate the development of more accurate PRSs.

In this manuscript, we develop a statistical method, termed as 'SummaryAUC', for approximating AUC and its variance for a given PRS based on summary statistics from an independent GWAS validation dataset. Although SummaryAUC relies on the normality assumption of PRS, extensive simulation studies demonstrate that it is highly accurate under realistic situations with more than five SNPs. Furthermore, SummaryAUC is flexible for PRSs with independent SNPs or SNPs in weak LD and for both genotyped and imputed SNPs in the validation dataset. Finally, we applied SummaryAUC to schizophrenia GWAS to demonstrate the validity of the methods. SummaryAUC is best used for PRSs with independent SNPs and for PRSs with <20 000 SNPs in weak LD for both accuracy and computational efficiency. SummaryAUC is not suitable for PRSs integrating all common SNPs in the genome, e.g. LD-Pred (Vilhjalmsson *et al.*, 2015) and BLUP-type PRSs (Golan and Rosset, 2014; Speed and Balding, 2014) based on linear mixed models. An R package with the same name was developed to implement the proposed method and is publicly available.

2 Materials and methods

We assume an additive PRS model based on M SNPs:

$$PRS_i = \sum_{m=1}^{M} w_m g_{im}, \tag{1}$$

where m indexes SNPs and i indexes subjects in the validation dataset. The weights $(w_1, \ldots, w_M)$ are derived based on a specific algorithm and a training dataset. The genotypic value $g_{im} \in \{0, 1, 2\}$ if the SNP is genotyped and $g_{im} \in [0,\ 2]$ if the SNP is imputed. The selected SNPs in the PRS may be correlated because of LD.

When the genotypic data and the binary phenotypic data (y_i) are available for each individual subject in the validation dataset, one can calculate PRS in (1) for all subjects and evaluate the performance of the prediction model by comparing PRS with the known phenotypic data. The performance of a prediction model is often assessed by the area under the AUC at the observational scale.

We are interested in developing methods for estimating AUC and its standard deviations when only the GWAS summary statistics are available for the validation dataset. For the mth SNP, the summary statistics include the minor allele, the minor allele frequency (MAF) in the control samples, the OR_m or equivalently the regression coefficient $\beta_m = \log(\mathrm{OR}_m)$, the two-sided P-value P_m or equivalently the Z-score statistic $Z_m = \mathrm{sign}(\beta_m)\Phi^{-1}(1 - P_m/2)$, the imputation information score r_m^2, the number of cases n_1 and the number of controls n_0. Here, β_m and P_m are based on single variant logistic regression. $\Phi()$ is the cumulative distribution function for $N(0,\ 1)$. In addition, MAF may not be available from the summary statistics to prevent subjects in the study to be deidentified (Homer *et al.*, 2008; Jacobs *et al.*, 2009).

2.1 Estimating AUC and its variance by summary statistics

Let $\mathrm{PRS}_{1i} = \sum_{m=1}^{M} w_m g_{im}^1$ be PRS for the i^{th}case and $\mathrm{PRS}_{0j} = \sum_{m=1}^{M} w_m g_{jm}^0$ for the jth control subject. We assume that M is reasonably large that PRSs approximately follow normal distributions in cases and controls, respectively:

$$PRS_{1i} \sim N(\mu_1, \sigma_1^2) \text{ and } PRS_{0j} \sim N\left(\mu_0, \sigma_0^2\right). \tag{2}$$

We will investigate the impact of the normality assumption in numerical studies. We define

$$\Delta = \frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}}. \tag{3}$$

The AUC (denoted as θ) is defined as the probability that, for a randomly selected case and a randomly selected control, the case has a larger PRS than the control, i.e. $\theta = P(PRS_{1i} > PRS_{0j})$. For a validation dataset with n_1 cases and n_0 controls with individual PRS values, one can estimate the AUC based on the following U-statistic (or the Wilcoxon–Mann–Whitney statistic):

$$\hat{\theta} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} I(PRS_{1i} > PRS_{0j}). \tag{4}$$

The expectation of AUC assuming normal distributions defined in (2) can be calculated as

$$\begin{aligned} \theta &= P(\mathrm{PRS}_{1i} > \ \mathrm{PRS}_{0j}) \\ &= P\left(\frac{(\mathrm{PRS}_{1i} > \ \mathrm{PRS}_{0j}) - (\mu_1 - \mu_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}} > \frac{-(\mu_1 - \mu_0)}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right) \\ &= \Phi(\Delta). \end{aligned} \tag{5}$$

When $n_1 \gg 1, \ n_0 \gg 1$ and $\sigma_1^2 \approx \sigma_0^2$ (when common SNPs have modest effect sizes for complex diseases), we derive in Supplementary Appendix that

$$\mathrm{Var}(\hat{\theta}) = \frac{n_1 + n_0}{n_1 n_0} \left(P(z_1 > -\Delta, \ z_2 > -\Delta) - \Phi^2(\Delta)\right), \tag{6}$$

where $z_1 \sim N(0,1)$, $z_2 \sim N(0,1)$ and $\mathrm{cor}(z_1, z_2) = 1/2$ If a PRS has rare SNPs with large effect size, variance is derived in Supplementary Appendix (A8) without assuming $\sigma_1^2 \approx \sigma_0^2$.

Note that both AUC (5) and the variance of the AUC estimator (6) depend on Δ defined in (3). Thus, it remains to estimate Δ based on the summary statistics in the validation dataset.

2.2 Estimate Δ when SNPs are independent

Let p_{1m} and p_{0m} denote the MAF of SNP m in the case and the control group, respectively. Typically, p_{0m} is included in the GWAS summary data while p_{1m} is not included. $p_{1m} = p_{0m}/(p_{0m}\mathrm{OR}_m + 1 - p_{0m})$, where OR_m is the OR in the validation dataset. Let $\tau_{0m}^2 = \mathrm{Var}(g_{jm}^0) = 2p_{0m}(1 - p_{0m})r_m^2$ be the genotypic variance in the control group assuming the Hardy Weinberg Equilibrium law. Similarly, let $\tau_{1m}^2 = \mathrm{Var}\left(g_{im}^1\right) = 2p_{1m}(1 - p_{1m})r_m^2$ be the genotypic variance in the case group. Remember that we assume $PRS_{1i} \sim N(\mu_1, \sigma_1^2)$ and $PRS_{0j} \sim N(\mu_0, \sigma_0^2)$ approximately in (2). When SNPs in the PRS are independent, we have $\mu_1 = \sum_{m=1}^{M} 2w_m p_{1m}$, $\mu_0 = \sum_{m=1}^{M} 2w_m p_{0m}$, $\sigma_1^2 = \sum_{m=1}^{M} w_m^2 \tau_{1m}^2$ and $\sigma_0^2 = \sum_{m=1}^{M} w_m^2 \tau_{0m}^2$.

Substituting these into (3) leads to

$$\Delta = \frac{\mu_1 - \mu_0}{\sqrt{\sigma_1^2 + \sigma_0^2}} = \frac{\sum_{m=1}^{M} 2w_m(p_{1m} - p_{0m})}{\sqrt{\sum_{m=1}^{M} w_m^2(\tau_{1m}^2 + \tau_{0m}^2)}}. \tag{7}$$

Thus, it remains to estimate $p_{1m} - p_{0m}$, the difference of the allele frequencies between cases and controls.

Let G_m^1 and G_m^0 be the average genotypic values in the case group and the control group, respectively. One can estimate $p_{1m} - p_{0m}$ as

$$\hat{p}_{1m} - \hat{p}_{0m} = (G_m^1 - G_m^0)/2. \tag{8}$$

The Z-statistic for genetic association can be approximated by the t-statistic for large studies, i.e. $Z_m = (G_m^1 - G_m^0)/\sqrt{\tau_{1m}^2/n_1 + \tau_{0m}^2/n_0}$; thus, we have

$$G_m^1 - G_m^0 \approx Z_m \sqrt{\tau_{1m}^2/n_1 + \tau_{0m}^2/n_0}. \tag{9}$$

Combining (7), (8) and (9) leads to an estimate for Δ:

$$\hat{\Delta} = \frac{\sum_{m=1}^{M} w_m Z_m \sqrt{\tau_{1m}^2/n_1 + \tau_{0m}^2/n_0}}{\sqrt{\sum_{m=1}^{M} w_m^2(\tau_{1m}^2 + \tau_{0m}^2)}}. \tag{10}$$

When p_{0m} is not included in the summary statistics, we can use the allele frequency based on public data (e.g. The 1000 Genome Project) of the similar ancestry populations to approximate τ_{0m}^2 and τ_{1m}^2.

2.3 Estimate Δ when SNPs are in LD

We first assume that SNPs in PRS are genotyped in the validation dataset. When some SNPs in PRS are in LD, we only need to modify the denominator in (10). We assume that, for complex diseases, correlations between local SNPs are similar in cases and controls. This assumption is reasonable because ORs are modest for nearly all disease-associated SNPs. Let $\rho_{ml} = \mathrm{cor}(g_{im}, g_{il})$ be the genotypic correlation between SNP m and SNP l. The variance of PRS in controls and cases are

$$\mathrm{Var}\left(PRS_{0j}\right) = \sum\nolimits_{m=1}^{M} w_m^2 \tau_{0m}^2 + 2\sum\nolimits_{m<l} w_m w_l \tau_{0m} \tau_{0l} \rho_{ml}$$

and

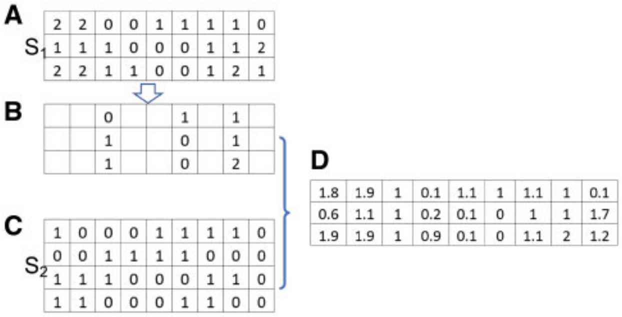
$$\mathrm{Var}(PRS_{1i}) = \sum\nolimits_{m=1}^{M} w_m^2 \tau_{1m}^2 + 2\sum\nolimits_{m<l} w_m w_l \tau_{1m} \tau_{1l} \rho_{ml}.$$

Thus, (10) can be modified as

$$\hat{\Delta} = \frac{\sum_{m=1}^{M} w_m Z_m \sqrt{\tau_{1m}^2/n_1 + \tau_{0m}^2/n_0}}{\sqrt{\sum_{m=1}^{M} w_m^2(\tau_{1m}^2 + \tau_{0m}^2) + 2\sum_{m<l} w_m w_l (\tau_{0m}\tau_{0l} + \tau_{1m}\tau_{1l})\rho_{ml}}}. \tag{11}$$

In implementation, we assume $\rho_{ml} = 0$ for SNPs located on different chromosomes and for SNPs on the same chromosome but more than 5 Mb away. In addition, we estimated ρ_{ml} using the genotype data of the similar ancestry population in The 1000 Genomes Project.

However, it is very challenging when SNPs in the PRS are imputed. Let $\hat{g}_{im}$be the imputed genotypic dosage. Let $\rho'_{ml} = \mathrm{cor}(\hat{g}_{im}, \hat{g}_{il})$ be the correlation between the imputed genotypic dosages. Although we cannot rigorously prove, we observe that imputation tends to inflate the magnitude of pairwise correlation, particularly for SNPs imputed with high uncertainty. We find that using ρ_{ml} (calculated based on genotype data in external data) to calculate $\hat{\Delta}$ in (11) makes the approximation to AUC and its variance less accurate, particularly when a PRS has many SNPs. To address this problem, we propose a strategy to estimate ρ'_{ml} using The 1000 Genome Project data, which is illustrated in Figure 1. Briefly, the subjects in The 1000 Genome Project with relevant ancestry are divided into two sets, denoted as S_1 and S_2. For subjects in S_1, we keep only SNPs that are genotyped in the validation GWAS dataset and perform imputation to derive genotypic dosages using the



2	2	0	0	1	1	1	1	0
1	1	1	0	0	0	1	1	2
2	2	1	1	0	0	1	2	1

		0			1		1	
		1			0		1	
		1			0		2	

1	0	0	0	1	1	1	1	0
0	0	1	1	1	1	0	0	0
1	1	1	0	0	0	1	1	0
1	1	0	0	0	1	1	0	0

1.8	1.9	1	0.1	1.1	1	1.1	1	0.1
0.6	1.1	1	0.2	0.1	0	1	1	1.7
1.9	1.9	1	0.9	0.1	0	1.1	2	1.2

Fig. 1. Estimate correlation of imputed SNPs in the validation GWAS dataset. The subjects in The 1000 Genome Project with relevant ancestry are divided as two sets S_1 and S_2. (**A**) The genotype of the subjects in S_1. (**B**) Only SNPs that are genotyped in the validation GWAS dataset are kept. (**C**) The haplotypes in S_2 are used as reference panel for imputation. (**D**) Imputation is performed to derive the genotypic dosage for SNPs that are not genotyped in the validation GWAS dataset. The correlation between two SNPs is calculated based on the imputed genotypic dosages

haplotypes in S_2 as the reference panel. We can calculate $\rho'_{ml} = \text{cor}(\hat{g}_{im}, \hat{g}_{il})$ using the imputed genotypic dosages to approximate $\hat{\Delta}$ in (11). Obviously, the degree of uncertainty of imputed genotypes is similar to that in the validation GWAS dataset because they are based on the same set of genotyped SNPs. Thus, we expect this strategy to attenuate the impact of imputation to the calculation of LD and thus to improve the approximation to AUC and its variance.

3 Results

3.1 Implementation

We implemented our algorithms in an R package 'SummaryAUC', which is freely available online. Pairwise correlations between SNPs are estimated using the genotype data in the The 1000 Genome Project with relevant ancestry. For PRS with practically independent SNPs (e.g. SNPs pruned rigorously), there is no limitation on the number of SNPs in PRS. When SNPs in a PRS are correlated, we only adjust for correlations for SNP pairs that are <5 Mb away. In real data analyses, most PRSs have <10 000 SNPs, which can be calculated in a few minutes.

3.2 Simulation studies

The key assumption of SummaryAUC is that PRS follows a normal distribution approximately. This assumption may lead to poor approximation to AUC when the number of SNPs (M) in a PRS is small. Thus, we performed simulations to investigate whether and how the accuracy of SummaryAUC depends on M. In each simulation, we simulated genotypes for 3000 cases and 3000 controls and for M independent SNPs ($M = 5,10,\cdots,100$). The allele frequency p_{0m} in controls followed a uniform distribution $U(0.05,\ 0.5)$. We simulated $\beta_m = \log(\text{OR}_m) \sim N(0,\ 1/3^2)$. The coefficients for PRS were set as the simulated β values. The allele frequency p_{0m} in cases were calculated as $p_{1m} = \text{OR}_m p_{om}/(\text{OR}_m p_{om} + 1 - p_{om})$, where OR_m is the OR for the SNP in the validation sample. Genotypic data were simulated using binomial distribution separately for cases and controls.

For each set of simulated data, we calculated AUC and its variance in two ways. In the first approach, we calculated PRS for each individual subject using the individual genotypic values; we then calculated AUC using an R package 'AUC' and its variance using bootstrap ($N = 10\,000$). In the second approach, we first performed association test for each SNP to derive Z_m and also allele frequency p_{0m} for control samples; we then approximated AUC and its variance using SummaryAUC.

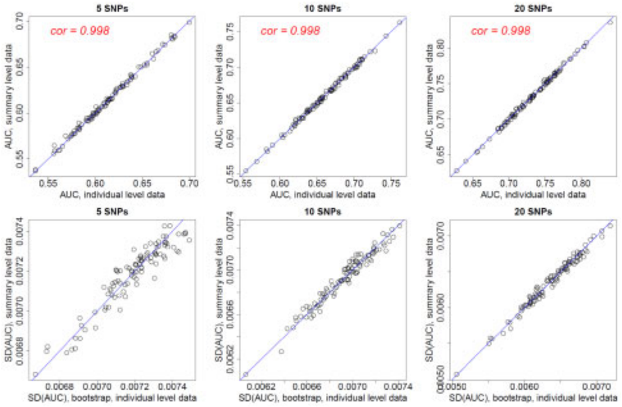


Fig. 2. AUC values and the standard errors for PRS with independent SNPs based on simulation study. Each data point represents one simulation. For each simulation, we calculated the AUC and its variance based on individual data (x-coordinate) and using SummaryAUC (y-coordinate)

The simulation results are summarized in Figure 2. Results suggest that SummaryAUC provides accurate approximation to AUC and its variance even when PRS has only five SNPs. Note that the correlation between true AUC and approximated AUC is >99%. Thus, the performance of SummaryAUC is robust to the number of SNPs in PRS.

Next, we also performed simulations by simulating $\beta_m = \log(\text{OR}_m) \sim N(0, 1/2^2),\ \sim N(0,1)$ and $\sim 0.25 + U(0, 0.5)$. The 90% quantile of the simulated ORs are 2.28, 5.2 and 2.01, respectively. These values are quite big for common SNPs and polygenic diseases. Results are reported in Supplementary Figures S1–S3. Again, SummaryAUC provides good approximations.

Finally, we performed simulations using SNPs that have been reported to be associated with nine complex diseases, including multiple autoimmune diseases and cancers (Supplementary Fig. S4A). The number of SNPs in PRSs ranged from 19 (melanoma) to 165 (prostate cancer). We used the published ORs and reference allele frequencies in the European ancestry for simulations. For most of the nine diseases, there was one common SNP with OR much bigger than the other SNPs (Supplementary Fig. S4B), providing an opportunity to check the robustness of SummaryAUC in presence of outliers in ORs. In all simulations, SummaryAUC provided accurate approximation to AUCs and their variances. Simulation results are summarized in Supplementary Figures S5 and S6.

3.3 Application to genetic risk prediction of schizophrenia

We evaluated the performance of SummaryAUC using schizophrenia GWAS (Schizophrenia Working Group of the Psychiatric Genomics, 2014). Schizophrenia is a devastating psychiatric disorder with high heritability (80–85%) and has a prevalence of ~1% worldwide. Schizophrenia is highly polygenic and estimated to be caused by more than 10 000 common SNPs. The Schizophrenia Working Group of Psychiatric Genetics Consortium (PGC) recently performed a meta-analysis of 49 case-control studies (34 241 cases and 45 604 controls) and 3 family studies (1235 parent affected offspring trios) and identified 108 genome-wide significant common SNPs. Encouragingly, PRSs have achieved a high discrimination performance with average AUC ~75% by leave-one-out analysis. Thus, this dataset is very useful for evaluating the performance of SummaryAUC because we can choose PRS with wide range of AUC values. Another advantage of using schizophrenia GWAS data is that the performance is typically maximized when most of SNPs are included in the PRS (Schizophrenia Working Group of the Psychiatric Genomics, 2014), i.e. the *P*-value threshold for including SNPs in PRS is nearly one. Thus, we can evaluate the accuracy of our method in presence of extensive correlations between SNPs.

PRSs were constructed using the results of a fixed-effect meta-analysis using all sub studies excluding the Molecular Schizophrenia Genetics (MGS) study (Shi *et al.*, 2009). The MGS study (2681 cases and 2653 controls of European ancestry) was used to calculate AUC values as the independent validation dataset. The meta-analysis results included *P*-values P_m and OR_m based in single variant logistic regression analysis. Let $w_m = \log(\mathrm{OR}_m)$, PRSs were built in two steps following Purcell's approach (Purcell *et al.*, 2009): (1) LD-clumping using pairwise correlation threshold *r*. LD-clumping was guided by the association *P*-values in the training dataset to keep the SNPs with smaller *P*-values in each specified short interval. After LD-clumping, reminding SNPs (denoted as A_r) had pairwise correlation less than *r*. (2) The PRS was defined as $PRS_i(Q, r) = \sum_{m \in A_r; P_m < Q} w_m g_{im}$ for a given *P*-value threshold *Q*. Note that the number of SNPs in PRS increases with *Q* and *r*. We chose a wide range of *P*-value threshold, ranging from 5×10^{-8} (genome-wide significance) to 1 (i.e. all SNPs in A_r). To investigate how the pairwise correlation in SNPs impacted the performance, we chose $r^2 = 0.01$ (very stringent, SNPs practically independent), 0.1 and 0.2 (locally, modestly correlated).

3.3.1 Compare performance when SNPs are genotyped in MGS

In the first set of analyses, we compared AUCs and SEs for PRSs restricted to genotyped SNPs in the MGS dataset. For each PRS, we calculated AUC and its standard error (SE) using MGS as the validation dataset in two ways. First, we assumed that individual level genotype/phenotype data in MGS were available, calculated PRS for each subject in MGS and calculated AUC using an R package 'AUC'. Then, we performed bootstrap ($N = 10\,000$) to estimate the SEs of the estimated AUC values. We denote the two values as AUC_0 and SE_0. Second, we performed single variant logistic regression to derive *P*-values, allele frequencies in controls and ORs for all common SNPs, adjusting for sex, age and the top 10 Principal component analysis (PCA) scores. The AUC and its SE were calculated using SummaryAUC. We denote the two values as AUC_1 and SE_1.

Results are reported in Figure 3. As was reported previously (Purcell *et al.*, 2009; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), including more SNPs in PRS increases AUC for schizophrenia because of the extremely highly polygenic genetic architecture. When SNPs are practically independent with pruning criteria $r^2 = 0.01$, AUC_0 and AUC_1 agree very well with the largest difference $|\mathrm{AUC}_0 - \mathrm{AUC}_1| = 0.37\%$. When we allow SNPs to be weakly correlated with pruning criteria $r^2 = 0.2$, we observed highly concordant results until *P*-value threshold <0.1, where the largest difference $|\mathrm{AUC}_0 - \mathrm{AUC}_1| = 0.31\%$. When we included SNPs with more liberal *P*-values, we observed larger inconsistency but the difference is still acceptable with the largest difference $|\mathrm{AUC}_0 - \mathrm{AUC}_1| = 0.63\%$ when all SNPs (122 552 SNPs) after pruning are included in PRS. Because we only ran 10 000 bootstrap samples to derive SE_0, there is some fluctuation across PRS models. Apparently, SE_1 provides an accurate approximation to SE_0.

To empirically examine the accuracy for PRSs with a small number of SNPs, we examined the performance of SummaryAUC for PRS with the number of SNPs varying from 5 to 100 (Fig. 4). When $r^2 = 0.01$, the largest difference $|\mathrm{AUC}_0\text{-}\mathrm{AUC}_1| = 0.41\%$; when $r^2 = 0.2$, the largest difference $|\mathrm{AUC}_0 - \mathrm{AUC}_1| = 0.55\%$. SE_1 values also agree well although SE_0 fluctuates because of the limited number of bootstrap samples.

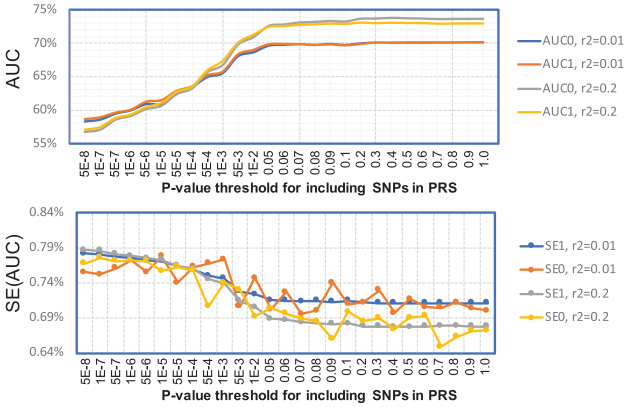


Fig. 3. AUC values and their standard errors for PRS of schizophrenia for genotyped SNPs in the MGS study. PRSs were trained based on the PGC summary data excluding the MGS study; the MGS study was used as the validation data for calculating AUC and its variance. Analysis was restricted to SNPs genotyped in MGS. The *x*-coordinate is the *P*-value threshold (for training dataset) for including SNPs in PRS. AUC_0 and SE_0 were calculated using individual level data. AUC_1 and SE_1 were calculated using SummaryAUC based on GWAS summary data in MGS. $r^2 = 0.01$: SNPs were LD-clumped using $r^2 = 0.01$ in PLINK. $r^2 = 0.2$: SNPs were LD-clumped using $r^2 = 0.2$ in PLINK

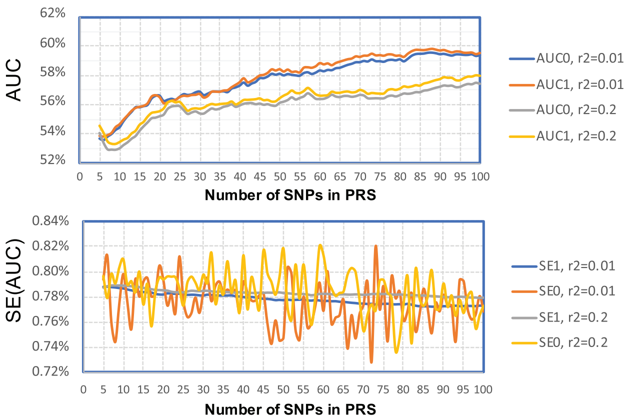


Fig. 4. AUC values and their standard errors for PRS of schizophrenia for genotyped SNPs in validation dataset. PRSs were trained based on the PGC data excluding the MGS study; the MGS study was used as the validation data for calculating AUC and its variance. The *x*-coordinate is the number of SNPs (with smallest *P*-values in the training dataset) for PRS. AUC_0 and SE_0 were calculated assuming individual level data. AUC_1 and SE_1 were calculated using SummaryAUC based on GWAS summary data in MGS. $r^2 = 0.01$: SNPs were LD-clumped using $r^2 = 0.01$ in PLINK. $r^2 = 0.2$: SNPs were LD-clumped using $r^2 = 0.2$ in PLINK

3.3.2 Compare performance when SNPs are imputed in MGS

MGS samples were imputed using software IMPUTE2 (Howie *et al.*, 2009) and using the haplotypes in The 1000 Genome Project as the reference. SNPs with imputation $R^2 < 0.5$ were excluded from analyses. Again, AUC_0 and SE_0 were calculated using individual level data. AUC_1 and SE_1 were calculated using SummaryAUC with $\rho_{ml} = \mathrm{cor}(g_{im},\ g_{il})$ estimated directly using the genotype data in The 1000 Genome Project. AUC_2 and SE_2 were calculated using

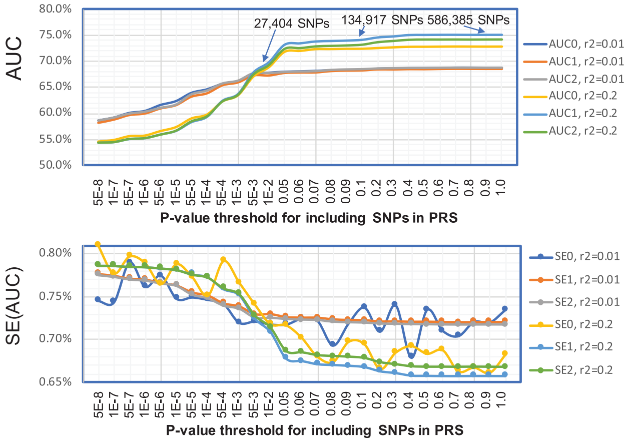


Fig. 5. AUC values and their standard errors for PRS of schizophrenia when SNPs are imputed. PRSs were trained based on the PGC data excluding the MGS study; the MGS study was used as the validation data for calculating AUC and its variance. MGS samples were imputed using IMPUTE2 and the haplotypes in The 1000 Genome Project. AUC_0 and SE_0 were calculated assuming individual level data. AUC_1 and SE_1 were calculated using SummaryAUC based on GWAS summary data in MGS, where pairwise SNP correlation was estimated using the genotype in The 1000 Genome Project directly. AUC_2 and SE_2 were calculated using SummaryAUC based on the summary data in MGS, where pairwise correlation was estimated using the imputed genotypic dosage data illustrated in Figure 1. $r^2 = 0.01$: SNPs were LD-clumped using $r^2 = 0.01$ in PLINK. $r^2 = 0.2$: SNPs were LD-clumped using $r^2 = 0.2$ in PLINK

SummaryAUC with $\rho'_{ml} = \text{cor}(\hat{g}_{im}, \hat{g}_{il})$ estimated using the imputed genotypic dosage data for samples in The 1000 Genome Project, as illustrated in Figure 1.

Results are presented in Figure 5. When SNPs were very rigorously pruned with $r^2 = 0.01$, both methods approximated AUCs and their standard errors very well. When SNPs were pruned with $r^2 = 0.2$, both AUC_1 and AUC_2 approximate AUC_0 very well when the *P*-value threshold <0.01. In fact, when *P*-value threshold =0.01, the PRS has 27 404 SNPs. When PRSs use more liberal *P*-value threshold to increase the number of SNPs, precisely adjusting for local correlation between SNPs becomes more important but difficult for SummaryAUC. In this case, AUC_1 does not approximate AUC_0 very well with the largest bias $|\text{AUC}_0 - \text{AUC}_1| = 2.25\%$. This is because AUC_1 uses $\rho_{ml} = \text{cor}(g_{im},\ g_{il})$ estimated directly using the genotype data in The 1000 Genome Project. In fact, imputation may change the correlation for imputed SNPs, particularly for poorly imputed SNPs. Encouragingly, AUC_2 better approximates AUC_0 with the largest bias $|\text{AUC}_0 - \text{AUC}_2| = 1.30\%$. The remaining bias may be due to the subtle difference of LD between the external genotype and the MGS population. A similar pattern is observed for SE_0, SE_1 and SE_2.

4 Discussions

Although the predictive performance of PRS models are relatively poor for most of complex diseases, PRS will be improved by increasing the sample size of the training GWAS dataset and innovative statistical methods that incorporate additional biological information, e.g. functional annotation data and genetic pleiotropy (T.Chen et al., submitted for publication; Hu *et al.*, 2017; Shi *et al.*, 2016). One difficulty for developing more accurate PRS is to evaluate the predictive performance of PRS in independent GWAS, which requires individual level genotypic and phenotypic data. Herein, we develop SummaryAUC as a new statistical method for approximating AUC and its standard error based on GWAS summary level data, which will greatly facilitate the development of more accurate PRS models.

Although SummaryAUC was derived under the normality assumption of PRS, simulation results suggest that the performance of SummaryAUC is robust to the number of SNPs in PRS. In fact, we found that SummaryAUC was accurate even for PRS with only five SNPs.

We systematically examined the accuracy of SummaryAUC by applying it to schizophrenia GWAS. Because of the extremely polygenic genetic architecture and the very large sample size in the training dataset, PRS can reach as high as 75% when all SNPs after LD-pruning are included. These data are very useful for examining the performance of SummaryAUC because we can compare accuracy for a wide range of AUC and for PRSs with tens of thousands of SNPs. The observations from this numerical study can be summarized as follows.

First, when SNPs in PRS are practically independent after rigorous LD-pruning, SummaryAUC is most accurate and the accuracy is not compromised even when all SNPs (after pruning) are included in PRS. In this case, computation is very fast.

Second, if locally correlated, genotyped SNPs are included in PRS, the performance is only slightly compromised compared to that based on independent SNPs. Because we have to adjust for local correlation, the computation might be slightly slow. In our implementation, we set $\rho_{ml} = 0$ for SNPs on different chromosomes or on the same chromosome but 5 Mb away; thus, computation can be done within a few minutes even when PRS has tens of thousands of SNPs.

Third, it is more complicated when PRS has correlated SNPs that are imputed in the validation GWAS dataset. When PRS is reasonably sparse (e.g. <20 000 SNPs), SummaryAUC is still accurate. However, when PRS has many SNPs (e.g. when including SNPs using *P*-value threshold 0.01), SummaryAUC is less accurate if pairwise correlations are not appropriately adjusted. In this case, an imputation-based method helps to reduce the bias. In reality, this only applies to schizophrenia and a few other psychiatric disorders because of their highly polygenic genetic architecture. For most of other diseases, the PRS with optimal classification accuracy is sparse, typically with <2000 SNPs. Thus, we expect SummaryAUC to work well.

Thus, if PRSs use independent SNPs, SummaryAUC can be used with confidence for PRS with any size and for both genotyped and imputed SNPs. If PRSs use correlated SNPs that are imputed in validation GWAS dataset, SummaryAUC is most accurate for sparse PRS models and needs to adjust correlations using imputed dosage data only for very dense PRS models. In addition, SummaryAUC is not suitable for PRSs using all common SNPs in the genome, e.g. LD-Pred (Vilhjalmsson *et al.*, 2015) and BLUP-type PRSs (Golan and Rosset, 2014; Speed and Balding, 2014) that are based on linear mixed models. It is computationally infeasible to adjust for the correlation for multiple millions of SNPs.

In addition, we found from simulations that, if summary statistics in the validation dataset were not appropriately corrected for population stratification, SummaryAUC may overestimate AUC. Thus, we recommend checking the quantile–quantile plot and running LD-score regression (Bulik-Sullivan *et al.*, 2015) to estimate the extent of population stratification. If population stratification is a major concern for a validation dataset (e.g. cases and controls are from different studies), it should not be used for evaluating AUC for a PRS. Finally, if the summary level data based on pooling multiple studies are used for validation, we need to consider heterogeneity (differences in ORs and allele frequencies for SNPs in PRS) across these studies. If evidence suggests strong and extensive heterogeneity across studies, the pooled dataset should not be used as the validation dataset and effort must be made to perform validation for individual studies.

Currently, we are working on developing methods for approximating $R^2 = \mathrm{cor}^2(y_i, PRS_i)$ using GWAS summary data in the validation dataset, the fraction of phenotypic variance explained by a PRS model at the observational scale. In addition, we are working on developing statistical methods for testing whether the AUC values from two PRS models are statistically different using GWAS summary data.

Acknowledgements

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov). The Molecular Genetics of Schizophrenia (MGS) Consortium includes P.V. Gejman, A.R. Sanders, J. Duan (North Shore University Health System and University of Chicago), C.R. Cloninger, D.M. Svrakic (Washington University, St Louis), N.G. Buccola (Louisiana State University Health Sciences Center, New Orleans), D.F. Levinson, J.S. (Stanford University, Stanford, California; Dr. J.S. is now at the National Cancer Institute), B.J. Mowry (Queensland Centre for Mental Health Research, Brisbane, and Queensland Brain Institute, University of Queensland, Brisbane), R. Freedman, A. Olincy (University of Colorado Denver), F. Amin (Atlanta Veterans Affairs Medical Center and Emory University, Atlanta), D.W. Black (University of Iowa Carver College of Medicine, Iowa City), J.M. Silverman (Mount Sinai School of Medicine, New York) and W.F. Byerley (University of California, San Francisco).

Funding

The project was supported by the NIH Intramural Research program.

Conflict of Interest: none declared.

References

Al Olama,A.A. *et al.* (2014) A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.*, **46**, 1103–1109.

Bulik-Sullivan,B.K. *et al.* (2015) LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.

Chatterjee,N. *et al.* (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.*, **45**, 400–405.

Chatterjee,N. *et al.* (2016) Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.*, **17**, 392–406.

Dudbridge,F. (2013) Power and predictive accuracy of polygenic risk scores. *PLoS Genet.*, **9**, e1003348.

Golan,D. and Rosset,S. (2014) Effective genetic-risk prediction using mixed models. *Am. J. Hum. Genet.*, **95**, 383–393.

Hanley,J.A. and Mcneil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (Roc) curve. *Radiology*, **143**, 29–36.

Homer,N. *et al.* (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.*, **4**, e1000167.

Howie,B.N. *et al.* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.

Hu,Y. *et al.* (2017) Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genet.*, **13**, e1006836.

Jacobs,K.B. *et al.* (2009) A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nat. Genet.*, **41**, 1253–1257.

Katki,H.A. and Schiffman,M. (2018) A novel metric that quantifies risk stratification for evaluating diagnostic tests: the example of evaluating cervical-cancer screening tests across populations. *Prev. Med.*, **110**, 100–105.

Kooperberg,C. *et al.* (2010) Risk prediction using genome-wide association studies. *Genet. Epidemiol.*, **34**, 643–652.

Krzanowski,W.J. and Hand,D.J. (2009) *ROC Curves for Continuous Data*. Chapman & Hall, CRC Press, London.

Maier,R. *et al.* (2015) Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.*, **96**, 283–294.

Michailidou,K. *et al.* (2017) Association analysis identifies 65 new breast cancer risk loci. *Nature*, **551**, 92–94.

Pepe,M.S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Purcell,S.M. *et al.* (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.

Scott,R.A. *et al.* (2017) An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes*, **66**, 2888–2902.

Shi,J. *et al.* (2009) Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*, **460**, 753–757.

Shi,J. *et al.* (2016) Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS Genet.*, **12**, e1006493.

Speed,D. and Balding,D.J. (2014) MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.*, **24**, 1550–1557.

Vilhjalmsson,B.J. *et al.* (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.*, **97**, 576–592.

Wei,Z. *et al.* (2009) From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet.*, **5**, e1000678.

Yang,J. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.

Zou,K.H. *et al.* (2011) *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*. Chapman and Hall, CRC.