Submitted to the Annals of Applied Statistics

ON PRS FOR COMPLEX POLYGENIC TRAIT PREDICTION

BY BINGXIN ZHAO AND FEI ZOU

University of North Carolina at Chapel Hill

Polygenic risk score (PRS) is the state-of-art prediction method for complex traits using summary level data from discovery genome-wide association studies (GWAS). The PRS, as its name suggests, is designed for polygenic traits by aggregating small genetic effects from a large number of causal SNPs and thus is viewed as a powerful method for predicting complex polygenic traits by the genetics community. However, one concern is that the prediction accuracy of PRS in practice remains low with little clinical utility, even for highly heritable traits. Another practical concern is whether genome-wide SNPs should be used in constructing PRS or not. To address the two concerns, we investigate PRS both empirically and theoretically. We show how the performance of PRS is influenced by the triplet (n, p, m) , where n, p, m are the sample size, the number of SNPs studied, and the number of true causal SNPs, respectively. For a given heritability, we find that i) when PRS is constructed with all p SNPs (referred as GWAS-PRS), its prediction accuracy is controlled by the p/n ratio; while ii) when PRS is built with a set of top-ranked SNPs that pass a pre-specified threshold (referred as threshold-PRS), its accuracy varies depending on how sparse the true genetic signals are. Only when m is magnitude smaller than n , or genetic signals are sparse, can threshold-PRS perform well and outperform GWAS-PRS. Our results demystify the low performance of PRS in predicting highly polygenic traits, which will greatly increase researchers' awareness of the power and limitations of PRS, and clear up some confusion on the clinical application of PRS.

1. Introduction. With the rapid development in biomedical technologies, various types of large-scale genetics and genomics data have been collected for better understanding of genetic etiologies underlying complex human diseases and traits. Genome-wide association studies (GWAS) aim to examine association between complex traits and single-nucleotide polymorphisms (SNPs). To detect SNPs that are associated with a given phenotype, GWAS commonly perform single SNP analysis to estimate and test the association between the phenotype and each candidate SNP one at a time,

Keywords and phrases: Polygenic risk score, Complex polygenic trait, High-dimensional prediction, Spurious correlation, Genetic risk score.

while adjusting for the effects of non-genetic factors and population sub-structures ([Price et al., 2006](#)). Tens of thousands of statistically significant SNPs have been detected for hundreds of human diseases/traits through GWAS ([MacArthur et al., 2016](#); [Visscher et al., 2017](#)). However, most of the identified SNPs have very low marginal genetic effects, explaining only a very small portion of the phenotypic variation even for traits with known high heritability ([Visscher et al., 2012](#)), resulting in a so called “missing heritability” phenomenon ([Manolio et al., 2009](#); [Zuk et al., 2012](#)).

One explanation for the missing heritability is that most complex traits are polygenic, affected by many genes whose individual effects are small ([Timpson et al., 2018](#)). The polygenicity of traits has long been hypothesized ([Penrose, 1953](#); [Fisher, 1919](#); [Gottesman and Shields, 1967](#)) and is supported by increasing empirical evidence ([Yang et al., 2015, 2010](#); [Dudbridge, 2016](#); [Ge et al., 2017](#); [Shi, Kichaev and Pasaniuc, 2016](#); [Lee et al., 2012](#); [Kemp et al., 2017](#); [Wray et al., 2018](#)). Given the polygenic nature of complex traits, it has been hypothesized that many causal SNPs are not likely to pass the genome-wide significance threshold, but should be informative and used for polygenic trait prediction. For these reasons, the polygenic risk score ([Purcell et al., 2009](#)) (PRS) is proposed, as a weighted sum of top-ranked candidate SNPs where each SNP is weighted by its estimated marginal effect from a discovery GWAS. As its name suggests, PRS aims to aggregate genetic effects of polygenes. Thus, it is expected to be powerful for polygenic and omnigenic traits ([Boyle, Li and Pritchard, 2017](#)). The omnigenic model is a newly emerging model for complex traits that are affected by the majority (if not all) of candidate SNPs.

In PRS, each SNP is weighted by its GWAS summary statistics, which avoids the need of accessing individual-level genotype data of discovery GWAS, and largely reduces the computational and data storage burdens and bypasses privacy concerns of sharing personal DNA information. As the GWAS summary statistics quickly accumulate in large publicly available data repositories ([Watanabe et al., 2018](#)), PRS has been widely used for complex traits in different domains. There were over 3,000 PRS-related publications in 2018. However, the prediction power of PRS remains disappointingly low with little clinical utility, even for traits with known high heritability ([Zheutlin and Ross, 2018](#); [Márquez-Luna, Loh and Price, 2017](#); [Torkamani, Wineinger and Topol, 2018](#)). Two legitimate reasons for the poor performance of PRS include 1) low quality SNP arrays with low coverage of causal SNPs; and 2) low quality top-ranked SNPs in tagging causal SNPs ([Chatterjee, Shi and García-Closas, 2016](#); [Wray et al., 2013](#)). However, as will be shown by the paper, even in the absence of the above two reasons,

POLYGENIC RISK SCORE PREDICTION

3

Figure 1: Impact of spurious correlation on the marginal SNP effect estimate of SNP i ($i = 1, \dots, p$).

PRS may still perform poorly. Thus far, except few studies (Daetwyler, Villanueva and Woolliams, 2008; Dudbridge, 2013; Chatterjee et al., 2013; Vilhjálmsson et al., 2015; Pasaniuc and Price, 2017), little research has been seriously done to study the statistical properties of PRS for polygenic and omnigenic traits.

We aim to fill the gap by empirically and theoretically studying PRS under various polygenic model assumptions with a hope to clear some misperceptions on PRS and to provide some practical guidelines on the use of PRS. Since PRS is built upon the marginal SNP effect estimates, we start our investigation from the statistical properties of the estimates. The performance of PRS is closely related to the asymptotic behavior of sure independent screening (Fan and Lv, 2008) (SIS) when signals are dense. It turns out, our theoretical investigation on the marginal genetic effect estimates is highly relevant to the “spurious correlation” problem (Fan, Guo and Hao, 2012) associated with high and ultra-high dimensional data, which provides another perspective on PRS. Linking the performance of single SNP analysis to the spurious correlation problem makes another significant contribution in understanding the behavior of the single SNP analysis, one of the most commonly used GWAS analysis approaches. For complex polygenic traits, spurious correlation makes the separation of causal and null SNPs difficult (Figure 1). As will be illustrated later, even for a fully heritable trait with a 100% genetic heritability, the estimated genetic effects of causal and null SNPs can be totally mixed and nonseparable from each other. The prediction power of PRS can go as low as zero.

In recognition of the relationship between the GWAS marginal screening and PRS, we show how the asymptotic prediction accuracy of PRS is affected by the triplet of (n, p, m) , where n is the sample size of the training

dataset, p is the total number of candidate SNPs, and m is the number of causal SNPs. Our investigation on PRS starts from the GWAS-PRS and is generalized to the threshold-PRS. Extensive simulations are carried out to evaluate PRS empirically and to evaluate our theoretical results under finite sample settings.

2. Single SNP Analysis. For a training dataset, let $\mathbf{y}$ be an $n \times 1$ phenotypic vector, $\mathbf{X}_{(1)}$ denote an $n \times m$ matrix of the causal SNPs and $\mathbf{X}_{(2)}$ denote an $n \times (p - m)$ matrix of the null SNPs, resulting in an $n \times p$ matrix of all SNPs denoted as $\mathbf{X} = [\mathbf{X}_{(1)}, \mathbf{X}_{(2)}] = [\mathbf{x}_1, \dots, \mathbf{x}_m, \mathbf{x}_{m+1}, \dots, \mathbf{x}_p]$. Columns of $\mathbf{X}$ are assumed to be independent for simplification. Furthermore, column-wise normalization is performed on $\mathbf{X}$ such that each SNP has mean zero and variance one. Define the following condition:

CONDITION 1. *Entries of $\mathbf{X} = [\mathbf{X}_{(1)}, \mathbf{X}_{(2)}]$ are real-valued independent random variables with mean zero, variance one, and a finite eighth order moment.*

The polygenic model assumes the following relationship between $\mathbf{y}$ and $\mathbf{X}$:

$$(1) \quad \mathbf{y} = \sum_{i=1}^p \mathbf{x}_i \beta_i + \boldsymbol{\epsilon} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m, \beta_{m+1}, \dots, \beta_p)^T$ is the vector of SNP effects such that the β_i s are i.i.d and follow $N(0, \sigma_\beta^2)$ for $i = 1, \dots, m$ and $\beta_i = 0$ for $i > m$. Let $\boldsymbol{\beta}_{(1)} = (\beta_1, \dots, \beta_m)^T$ and $\boldsymbol{\beta}_{(2)}$ be a $(p - m) \times 1$ vector with all elements being zero, and $\boldsymbol{\epsilon}$ represents the random error vector. For simplicity and without loss of generality, we assume that there exist no other covariates. According to the above model, the overall genetic heritability h^2 of $\mathbf{y}$ is

$$h^2 = \frac{\text{Var}(\mathbf{X}_{(1)} \boldsymbol{\beta}_{(1)})}{\text{Var}(\mathbf{y})} = \frac{\text{Var}(\mathbf{X}_{(1)} \boldsymbol{\beta}_{(1)})}{\text{Var}(\mathbf{X}_{(1)} \boldsymbol{\beta}_{(1)}) + \text{Var}(\boldsymbol{\epsilon})}.$$

For the rest of the paper, we set $h^2 = 1$, reducing the above model to the following deterministic model

$$(2) \quad \mathbf{y} = \sum_{i=1}^p \mathbf{x}_i \beta_i = \sum_{i=1}^m \mathbf{x}_i \beta_i = \mathbf{X}_{(1)} \boldsymbol{\beta}_{(1)},$$

the most optimistic situation in predicting phenotypes.

Though for a typical large-scale discovery GWAS, the sample size n is often not small (e.g., $n \sim 10,000$ or $100,000$), the number of candidate SNPs p is typically even larger (e.g., $p \sim 500,000$ or $10,000,000$). On the other hand, depending on their underlying genetic architectures, the number of causal SNPs can vary dramatically from one trait to another. To cover most of modern GWAS data, we therefore assume $n, p \rightarrow \infty$ and that

$$\frac{p}{n} = \gamma \rightarrow \gamma_0 \quad \text{and} \quad \frac{m}{p} = \omega \rightarrow \omega_0,$$

where $0 < \gamma_0 \leq \infty$, and $0 \leq \omega_0 \leq 1$.

For continuous traits in absence of any covariates, a typical single SNP analysis employs the following simple linear regression model

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{x}_i \beta_i + \boldsymbol{\epsilon}^*$$

for a given SNP i , where $\mathbf{1}_n$ is an $n \times 1$ vector of ones, β_i is its effect ($i = 1, \dots, p$). When both $\mathbf{y}$ and $\mathbf{x}_i$ are normalized and $n \rightarrow \infty$, under Condition (1) and polygenic model (2), the maximum likelihood estimate (MLE) of μ , $\hat{\mu} \equiv 0$, and the MLE of the genetic effect β_i equals

$$\hat{\beta}_i = (\mathbf{x}_i^T \mathbf{x}_i)^{-1} \mathbf{x}_i^T \mathbf{y} = \frac{1}{n} \mathbf{x}_i^T \mathbf{y} = \sum_{j=1}^m r_{ij} \beta_j,$$

where $r_{ij} = \mathbf{x}_i^T \mathbf{x}_j / n = \sum_{k=1}^n x_{ik} x_{jk} / n$ is the sample correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$, $j = 1, \dots, p$. Specifically, for SNP i , $i = 1, \dots, p$, we have

$$\hat{\beta}_i = \begin{cases} \beta_i + \sum_{j \neq i}^m r_{ij} \beta_j, & \text{for } i \in [1, m]; \\ \sum_{j=1}^m r_{ij} \beta_j, & \text{for } i \in [m+1, p]. \end{cases}$$

Given that SNPs in $\mathbf{X}$ are independent of each other, or the correlation ρ_{ij} for any ij th SNP pair ($i \neq j$) is zero, it can be shown that asymptotically $\hat{\beta}_i$ is an unbiased estimator of β_i such that

$$\mathbb{E}(\hat{\beta}_i) = \begin{cases} \beta_i, & \text{for } i \in [1, m]; \\ 0, & \text{for } i \in [m+1, p] \end{cases}$$

as $n \rightarrow \infty$. The associated variance of $\hat{\beta}_i$ grows linearly with m since for any causal SNP i ($1 \leq i \leq m$), we have

$$\begin{aligned} \text{Var}(\hat{\beta}_i) &= \text{Var}\left(\sum_{j \neq i}^m r_{ij} \beta_j\right) = \frac{1}{n^2} \sum_{j \neq i}^m \beta_j^2 \cdot \mathbb{E}\left(\sum_{k=1}^n x_{ik}^2 x_{jk}^2\right) \\ &= \frac{\sum_{j \neq i}^m \beta_j^2}{n} = \mathcal{O}\left(\frac{m}{n}\right) = \mathcal{O}(\gamma \cdot \omega). \end{aligned}$$

Similarly, for any null SNP i ($m < i \leq p$), we have

$$\text{Var}(\hat{\beta}_i) = \text{Var}\left(\sum_{j=1}^m r_{ij}\beta_j\right) = \frac{\sum_{j=1}^m \beta_j^2}{n} = \mathcal{O}\left(\frac{m}{n}\right) = \mathcal{O}(\gamma \cdot \omega).$$

Therefore, $\text{Var}(\hat{\beta}_i) = \mathcal{O}(m/n) = \mathcal{O}(\gamma \cdot \omega)$ for all i ($i = 1, \dots, p$), suggesting that the associated variances of all $\hat{\beta}_i$ s are in the same scale regardless of whether their corresponding SNPs are causal or not. When $m/n = \gamma \cdot \omega \rightarrow \gamma_0 \cdot \omega_0$ is large, $\hat{\beta}_i$ is no longer a reliable estimate of β_i . Moreover, the $\hat{\beta}_i$ s of the causal and null SNPs become well mixed and cannot be easily separated when m/n is large, which raises two important concerns include 1) how does this affect the SNP selection; and 2) how does this affect the weights in PRS which ultimately affect the performance of PRS? We address these two concerns below.

3. PRS. For a testing dataset with n_z samples, define its $n_z \times p$ SNP matrix as $\mathbf{Z} = [\mathbf{Z}_{(1)}, \mathbf{Z}_{(2)}]$ with $\mathbf{Z}_{(1)} = [\mathbf{z}_1, \dots, \mathbf{z}_m]$ and $\mathbf{Z}_{(2)} = [\mathbf{z}_{m+1}, \dots, \mathbf{z}_p]$. Then $\mathbf{y}_z = \sum_{i=1}^p \mathbf{z}_i \beta_i = \mathbf{Z}_{(1)} \boldsymbol{\beta}_{(1)}$ and the PRS equals

$$\hat{\mathbf{y}}_P = \sum_{i=1}^p \mathbf{z}_i \hat{d}_i = \mathbf{Z} \hat{\mathbf{d}} = \mathbf{Z}_{(1)} \hat{\mathbf{d}}_{(1)} + \mathbf{Z}_{(2)} \hat{\mathbf{d}}_{(2)},$$

where $\hat{\mathbf{d}} = (\hat{d}_1, \dots, \hat{d}_m, \hat{d}_{m+1}, \dots, \hat{d}_p)^T = (\hat{\mathbf{d}}_{(1)}^T, \hat{\mathbf{d}}_{(2)}^T)^T$, $\hat{d}_i = \hat{\beta}_i \cdot \mathbf{I}(|\hat{\beta}_i| > c)$, $\mathbf{I}(\cdot)$ is the indicator function, and c is a given threshold for screening SNPs. When $c = 0$, all candidate SNPs are used, leading to GWAS-PRS. The prediction accuracy of PRS is typically measured by

$$A_P = \frac{\mathbf{y}_z^T \hat{\mathbf{y}}_P}{\|\mathbf{y}_z\| \cdot \|\hat{\mathbf{y}}_P\|} = \frac{(\mathbf{Z}_{(1)} \boldsymbol{\beta}_{(1)})^T (\mathbf{Z}_{(1)} \hat{\mathbf{d}}_{(1)} + \mathbf{Z}_{(2)} \hat{\mathbf{d}}_{(2)})}{\|\mathbf{Z}_{(1)} \boldsymbol{\beta}_{(1)}\| \cdot \|\mathbf{Z}_{(1)} \hat{\mathbf{d}}_{(1)} + \mathbf{Z}_{(2)} \hat{\mathbf{d}}_{(2)}\|},$$

where $\|\cdot\|$ is the l_2 norm of a vector.

3.1. GWAS-PRS. In GWAS-PRS, $\hat{\mathbf{d}}_{(1)} = \hat{\boldsymbol{\beta}}_{(1)}$ and $\hat{\mathbf{d}}_{(2)} = \hat{\boldsymbol{\beta}}_{(2)}$. For simplification, in rest of the paper, we set $n_z = n$ and our general conclusions should remain the same when the two are different. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_m, \hat{\beta}_{m+1}, \dots, \hat{\beta}_p)^T = (\hat{\boldsymbol{\beta}}_{(1)}^T, \hat{\boldsymbol{\beta}}_{(2)}^T)^T$, where $\hat{\boldsymbol{\beta}}_{(1)} = \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \boldsymbol{\beta}_{(1)} / n$, and $\hat{\boldsymbol{\beta}}_{(2)} = \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \boldsymbol{\beta}_{(1)} / n$. Then

$$A_P = \frac{\boldsymbol{\beta}_{(1)}^T \mathbf{Z}_{(1)}^T (\mathbf{Z}_{(1)} \hat{\boldsymbol{\beta}}_{(1)} + \mathbf{Z}_{(2)} \hat{\boldsymbol{\beta}}_{(2)})}{\|\mathbf{Z}_{(1)} \boldsymbol{\beta}_{(1)}\| \cdot \|\mathbf{Z}_{(1)} \hat{\boldsymbol{\beta}}_{(1)} + \mathbf{Z}_{(2)} \hat{\boldsymbol{\beta}}_{(2)}\|} = \frac{C_1}{\{\text{Var}_1\}^{1/2} \cdot \{\text{Var}_2\}^{1/2}},$$

POLYGENIC RISK SCORE PREDICTION

7

Figure 2: Prediction accuracy (A_P) of GWAS-PRS across different m/p ratios when $c = 0$ (i.e., all SNPs are selected). We set $p=100,000$ and $n=100, 1000$ and $10,000$, respectively.

where

$$\begin{aligned} C_1 &= \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}, \\ \text{Var}_1 &= \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}, \quad \text{and} \\ \text{Var}_2 &= (\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T) \cdot \\ &\quad (\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}). \end{aligned}$$

THEOREM 1. Under the polygenic model (2) and Condition (1), if $m \rightarrow \infty$, $p \rightarrow \infty$, and $p/n^2 \rightarrow 0$ as $n \rightarrow \infty$,

$$(3) \quad A_P^2 / \left(\frac{n}{n+p} \right) = A_P^2 / \left(\frac{1}{1+\gamma} \right) = 1 + o_p(1).$$

If we further assume $p = c \cdot n^\alpha$ for some constants $c \in (0, \infty)$ and $\alpha \in (0, \infty]$, then

$$A_P^2 = \begin{cases} 1 + o_p(1), & \text{if } 0 < \alpha < 1; \\ 1/(1+c) + o_p(1), & \text{if } \alpha = 1; \\ o_p(1), & \text{if } \alpha > 1, \end{cases}$$

as $n \rightarrow \infty$.

Proof of Theorem 1 is given in Appendix A.

REMARK 1. A_P^2 has nonzero asymptotic limit provided that $\alpha \in (0, 1]$. As illustrated in Figure 2, A_P converges to zero if $\alpha \in [2, \infty]$, indicating the null prediction power of GWAS-PRS even for traits that are fully heritable.

REMARK 2. *Since causal SNPs are not known as a prior but estimated, a poorly selected SNP list can be a reason for the poor performance of PRS. However, Theorem 1 suggests that when m is large, even in the oracle situation where the selected SNP list contains all and only all of the causal SNPs, its performance is limited by the ratio of m/n . For complex traits underlying the omnigenic model (Boyle, Li and Pritchard, 2017), the expected prediction power of the oracle PRS is essentially zero.*

REMARK 3. *In our investigation, we set $h^2 = 1$ to reflect the most optimistic situation for phenotype prediction. The result in (3) can be easily extended to $0 < h^2 < 1$ as follows:*

$$A_P^2 / \left(\frac{n \cdot h^2}{n + p/h^2} \right) = 1 + o_p(1).$$

3.2. *Illustration of asymptotic limits of GWAS-PRS.* We numerically evaluate the analytical results above and the performance of A_P^2 with $p = 100,000$, and $n = n_z = 100, 1000$ or $10,000$. Each entry of $\mathbf{X}$ and $\mathbf{Z}$ is independently generated from $N(0, 1)$. We also vary the ratio of causal SNPs m/p from 0.01 to 1 to reflect a wide range of SNP signals, from very sparse to very dense situations, respectively. The effects of the causal SNPs $\beta_{(1)} \sim MVN(0, \mathbf{I}_m)$, based on which the phenotypes $\mathbf{y}$ and $\mathbf{y}_z$ are generated from model (2), where $\mathbf{I}_m$ is an $m \times m$ identity matrix. A total of 100 replications are conducted for each simulation setup.

Figure 2 shows the A_P values across different simulation setups. As expected, for GWAS-PRS, A_P remains nearly constant regardless of m and is close to $\sqrt{n/(n+p)}$. For small n , the A_P values are fairly close to zero with a large variance.

3.3. *Threshold-PRS.* As shown in Theorem 1, the asymptotic limit of A_P^2 associated with GWAS-PRS does not depend on m , the number of causal SNPs, but n , the sample size of the training dataset. Where sample size is much smaller than the number of candidate SNPs, GWAS-PRS has low prediction accuracy. It is thus natural for us to turn around and ask with a properly selected threshold c , whether the performance of PRS can be improved or not.

3.3.1. *General Setup.* For a given threshold $c > 0$, let $q = p \cdot a$ ($a \in (0, 1]$), where q is the number of selected SNPs, among which there are q_1 true causal SNPs and the remaining q_2 are null SNPs, that is, $q = q_1 + q_2$. Let $\mathbf{Z}_{(1)} = [\mathbf{Z}_{(11)}, \mathbf{Z}_{(12)}]$, $\mathbf{Z}_{(2)} = [\mathbf{Z}_{(21)}, \mathbf{Z}_{(22)}]$, $\mathbf{X}_{(1)} = [\mathbf{X}_{(11)}, \mathbf{X}_{(12)}]$, $\mathbf{X}_{(2)} =$

$[\mathbf{X}_{(21)}, \mathbf{X}_{(22)}]$, $\hat{\beta}_{(1)}^T = (\hat{\beta}_{(11)}^T, \hat{\beta}_{(12)}^T)$, and $\hat{\beta}_{(2)}^T = (\hat{\beta}_{(21)}^T, \hat{\beta}_{(22)}^T)$, where $\mathbf{Z}_{(11)}$, $\mathbf{X}_{(11)}$, and $\hat{\beta}_{(11)}$ correspond to the selected q_1 causal SNPs, and $\mathbf{Z}_{(21)}$, $\mathbf{X}_{(21)}$, and $\hat{\beta}_{(21)}$ correspond to the selected q_2 null SNPs. The prediction accuracy of threshold-PRS is therefore

$$A_P = \frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T (\mathbf{Z}_{(11)} \hat{\beta}_{(11)} + \mathbf{Z}_{(21)} \hat{\beta}_{(21)})}{\|\mathbf{Z}_{(1)} \beta_{(1)}\| \cdot \|\mathbf{Z}_{(11)} \hat{\beta}_{(11)} + \mathbf{Z}_{(21)} \hat{\beta}_{(21)}\|} = \frac{C_1}{\{\text{Var}_1\}^{1/2} \cdot \{\text{Var}_2\}^{1/2}},$$

where

$$\begin{aligned} C_1 &= \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)}, \\ \text{Var}_1 &= \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}, \quad \text{and} \\ \text{Var}_2 &= (\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T) \cdot \\ &\quad (\mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)}). \end{aligned}$$

THEOREM 2. *Under the polygenic model (2) and Condition (1), if $m, q_1, q_2 \rightarrow \infty$ as $\min(n, p) \rightarrow \infty$, and further if $\{m^2(q_1 + q_2)\}/(n^2 q_1^2) \rightarrow 0$, then we have*

$$A_P^2 / \left(\frac{nq_1^2}{nmq_1 + qm^2} \right) = 1 + o_p(1).$$

However, if $\{m^2(q_1 + q_2)\}/(n^2 q_1^2) \not\rightarrow 0$ as $\min(n, p) \rightarrow \infty$, then

$$A_P^2 = \mathcal{O}_p(n^{-1}) = o_p(1).$$

Proof of Theorem 2 is given in Appendix A. Theorem 2 shows that given n and m , A_P is determined by q_1 , the number of selected causal SNPs, and q , the total number of selected SNPs. Expressing q_1 as a function of q , or $q_1 = \phi(q)$, then A_P can be re-expressed as a function of q :

$$A_P^2(q) = \frac{n\phi(q)^2}{nm\phi(q) + qm^2}.$$

3.3.2. Role of $\phi(q)$. The function $\phi(q)$ is a non-decreasing function of q and plays an important role in determining A_P . The exact form of $\phi(q)$ is trait-dependent and not easy to obtain. But for the following two special examples, we can evaluate $\phi(q)$ straightforwardly and investigate its impact on A_P . We first study the marginal distribution of the $\hat{\beta}_i$ s, which is a mixture of two distributions, one corresponding to the causal SNP set and one to the null SNP set. Let $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_m, \hat{\beta}_{m+1}, \dots, \hat{\beta}_p)^T = (\hat{\beta}_{(1)}^T, \hat{\beta}_{(2)}^T)^T$. Given

Figure 3: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. We set $p=100,000$ and $n = 10,000$ in training data and $n=1000$ in testing data.

that β_i s in $\beta_{(1)}$ are i.i.d and follow $N(0, \sigma_\beta^2)$, and the remaining ones in $\beta_{(2)}$ are all zeros, it follows from the central limit theorem that

$$\hat{\beta} \sim \begin{cases} N(0, \sigma_\beta^2 \cdot \frac{n+m}{n}), & \text{for causal SNPs;} \\ N(0, \sigma_\beta^2 \cdot \frac{m}{n}), & \text{for non-causal SNPs.} \end{cases}$$

When $m/n = \gamma \cdot \omega \rightarrow \gamma_0 \cdot \omega_0 = 0$, the spread of the marginal distribution of the causal SNPs is much wider than that of the marginal distribution of the null SNPs, making the two distributions separable and single SNP analysis powerful. However, as the genetic signal gets denser and denser (or as m increases), the difference between the two distributions gets smaller and smaller, leading to two well mixed distributions and poorly performed single SNP analysis. To see how the ratio of m/n impacts single SNP analysis, we first approximate the density of $\hat{\beta}_c$ by

$$\begin{aligned} f(b) &\approx \frac{m}{p} \cdot N(0, \sigma_\beta^2 \cdot \frac{n+m}{n}) + \frac{p-m}{p} \cdot N(0, \sigma_\beta^2 \cdot \frac{m}{n}) \\ &= \frac{m}{p} \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp(-\frac{b^2}{2\sigma_1^2}) + \frac{p-m}{p} \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp(-\frac{b^2}{2\sigma_2^2}) \end{aligned}$$

with cumulative distribution function (CDF) given by

$$F(b) \approx \frac{m}{p} \cdot \Phi\left(\frac{b}{\sigma_1}\right) + \frac{p-m}{p} \cdot \Phi\left(\frac{b}{\sigma_2}\right),$$

where $\sigma_1^2 = \sigma_\beta^2 \cdot (1+m/n)$, $\sigma_2^2 = \sigma_\beta^2 \cdot m/n$, and $\Phi(x) = \int_{-\infty}^x \exp(-t^2/2)dt/\sqrt{2\pi}$ is the CDF of the standard normal random variable. Since the mixture distribution is symmetric about zero, without loss of generality, we consider one-sided test and SNPs with the largest $(100 \times a)\%$ estimated genetic effects ($0 < a < 1/2$) are selected. For a causal SNP, its selection probability κ_1 equals to

$$\begin{aligned} & \Pr[b > F^{-1}(1-a)|b \sim N(0, \sigma_1^2)] \\ &= 1 - \Pr[b \leq F^{-1}(1-a)|b \sim N(0, \sigma_1^2)] \\ &= 1 - \Phi\left[\frac{F^{-1}(1-a)}{\sigma_1}\right] = 1 - \Phi\left[\frac{F^{-1}(1-a)}{\sigma_\beta} \sqrt{\frac{n}{n+m}}\right]. \end{aligned}$$

Similarly for a given null SNP, its selection probability κ_2 is

$$\begin{aligned} & \Pr[b > F^{-1}(1-a)|b \sim N(0, \sigma_2^2)] \\ &= 1 - \Pr[b \leq F^{-1}(1-a)|b \sim N(0, \sigma_2^2)] \\ &= 1 - \Phi\left[\frac{F^{-1}(1-a)}{\sigma_2}\right] = 1 - \Phi\left[\frac{F^{-1}(1-a)}{\sigma_\beta} \sqrt{\frac{n}{m}}\right]. \end{aligned}$$

Therefore, among $q = p \cdot a = q_1 + q_2$ selected SNPs, we have

$$\begin{aligned} m \cdot \kappa_1 &= m \cdot \left[1 - \Phi\left(\frac{F^{-1}(1-a)}{\sigma_\beta} \sqrt{\frac{n}{n+m}}\right)\right] \quad \text{and} \\ (p-m) \cdot \kappa_2 &= (p-m) \cdot \left[1 - \Phi\left(\frac{F^{-1}(1-a)}{\sigma_\beta} \sqrt{\frac{n}{m}}\right)\right] \end{aligned}$$

causal and null SNPs, respectively. For a given a or equivalently c , $F^{-1}(1-a)/\sigma_\beta$ is the same for both causal and null SNPs. Therefore, the quality of top-ranked SNP list is largely determined by m/n . Remark 4 discusses the upper bounds of A_p under two extreme cases.

REMARK 4. When $n/m = o(1)$, it is easy to see $\kappa_1 = \kappa_2 \cdot \{1 + o(1)\}$. Therefore, when both q_1 and q_2 are large, $q_1/q \approx m/p$, and thus $q_1 = \phi(q) \approx \frac{m}{p} \cdot q$, yielding

$$(4) \quad A_P^2(q) = \frac{n\phi(q)^2}{nm\phi(q) + qm^2} \approx \frac{n}{np + p^2} \cdot q.$$

Therefore, A_P reaches its upper bound $\sqrt{n/(n+p)}$ at $q = p$, suggesting that the best performing PRS is the one without SNP selection or GWAS-PRS when the genetic signals are dense. On the other hand, when $m/n = o(1)$, κ_1 becomes much larger than κ_2 . Thus, causal SNPs can be relatively easy to detect by using single SNP analysis. As a increases, q_1 eventually gets saturated at m , and threshold-PRS reaches its upper performance limit $\sqrt{n/(n+m)}$ with $q = q_1 = \phi(q) = m$, the oracle case described in Remark 2.

In conclusion, the above analysis provides some practical guidelines on the use of PRS: 1) SNP screening should be avoided for highly polygenic or omnigenic traits where the m/n ratio is large; and 2) for monogenic and oligogenic traits (Timpson et al., 2018) with a small m/n ratio, threshold-PRS should be used.

4. More simulation studies.

4.1. *Threshold-PRS.* To illustrate the finite sample pattern of threshold-PRS, we simulate $p = 100,000$ uncorrelated SNPs. As in Figure 2, we (naively) generate SNPs from $N(0, 1)$. To study the effect of m/p , we vary the number of causal SNPs m and set it to 100, 1000, 10,000 and 50,000. The casual SNP effects $\beta_{(1)} \sim MVN(0, \mathbf{I}_m)$. The linear polygenic model (2) is used to generate phenotype $\mathbf{y}$. The sample size is set to 1000 and 10,000 for training data, and 1000 for testing data. For threshold-PRS, we consider a series of P -value thresholds $\{1, 0.8, 0.5, 0.4, 0.3, 0.2, 0.1, 0.08, 0.05, 0.02, 0.01, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$ (Márquez-Luna, Loh and Price, 2017). We name this simulation setting Case 1. Again, a total of 100 replications are conducted for each simulation condition.

Figure 3 and Supplementary Figure 1 display the performance of threshold-PRS across a series of m/p ratios in Case 1. As expected, the performance of GWAS-PRS (i.e. at P -value threshold of 1) stays nearly constant around $\sqrt{n/(n+p)}$ regardless of the m/p ratios, which is about 0.3 (shown in Figure 3) for n of 10,000 and 0.1 (shown in Supplementary Figure 1) for n of 1000. In contrast, the performance of threshold-PRS varies with m . When m is small compared to n , threshold-PRS performs significantly better than GWAS-PRS provided a reasonable c is chosen, which in general is small as shown by Supplementary Figure 1. For example, when $m = 100$ and $n = 1000$, threshold-PRS achieves its best performance at $c = 10^{-5}$, with A_P of 0.75, in contrast to its oracle performance which is about 0.95. Figure 3 shows that when m gets close to n or larger than n , the performance of threshold-PRS drops significantly regardless of c . When m gets close to n , its performance remains similar for a wide range of c values; and when

m gets much larger than n , its performance improves as c increases, and eventually reaches the performance level of GWAS-PRS.

In addition, we vary Case 1 settings to check the sensitivity of our results. In Case 2, we generate actual SNP genotype data where the minor allele frequency (MAF) of each SNP, f , is independently generated from Uniform $[0.05, 0.45]$ and SNP genotypes are independently sampled from $\{0, 1, 2\}$ with probabilities $\{(1-f)^2, 2f(1-f), f^2\}$, respectively according to the Hardy-Weinberg equilibrium principle. In Case 3, we simulate mixed samples from five subpopulations. The overall MAF of each SNP in mixed samples is then correspondingly generated from Uniform $[0.05, 0.45]$, and the F_{st} values are independently generated from Uniform $[0.01, 0.04]$ (Lee, Wright and Zou, 2011), based on which the MAF of each sub-population is generated according to the Balding-Nichols model (Balding and Nichols, 1995). We let the sample size of sub-population be the same, and set them at either 200 or 2000. The population substructures are estimated with the principal component analysis (Price et al., 2006) and the top 4 principal components are included as covariates in the single SNP analysis. Case 4 allows larger variability in the causal SNP effects such that β_i s are independently generated from $N(0, \sigma_i^2)$, where σ_i^{-2} follows a gamma distribution with $\alpha = 10$ and $\beta = 9$.

The results of Case 2 are displayed in Supplementary Figures 2 and 3, which are similar to those of Case 1. Supplementary Figures 4 and 5 display the oracle performance of PRS under varying m/p ratios in Case 2. Clearly A_p is around $\sqrt{n/(n+m)}$, confirming the poor performance of PRS even in the oracle case when genetic signals are dense. The performance of threshold-PRS under Case 2 for traits with $h^2=0.5$ is presented in Supplementary Figures 6 and 7. Compared to Supplementary Figures 2 and 3 where $h^2=1$, the prediction accuracy of threshold-PRS decreases, but the general patterns remain the same. The results of Case 3 are displayed in Supplementary Figures 8 and 9. In the presence of population substructures, if they are properly adjusted, the main pattern of threshold-PRS remains unchanged and the performance of GWAS-PRS agrees well with the theoretical results. The results of Case 4 are displayed in Supplementary Figures 10 and 11, which are also similar to those of Case 1, indicating that our asymptotic results are not sensitive to the distribution of the causal SNP effects, or $\beta_{(1)}$.

4.2. UK Biobank SNP data. We perform additional simulation based on real SNP data from the UK Biobank (UKB) resources (Sudlow et al., 2015; Bycroft et al., 2018). We download the unimputed genotype data released in

July 2017 and apply the following quality control (QC) procedures: excluding subjects with more than 10% missing genotypes, only including SNPs with $MAF > 0.01$, genotyping rate $> 90\%$, and passing Hardy-Weinberg test (P -value $> 1 \times 10^{-7}$). There are 461,488 SNPs left for 488,371 subjects after QC, and we randomly select 11,000 individuals of British ancestry to perform this simulation, among which 10,000 are randomly selected and used as training data, and PRS are constructed on the remaining 1,000 individuals. Causal SNPs are randomly selected, and the number of causal SNPs m is set to 470, 4700, 47,000, and 235,000. The nonzero SNP effects are independently generated from $N(0, 1)$, and the heritability h^2 is set to 80%. Single SNP analysis is performed using the PLINK toolset (Purcell et al., 2007). To obtain a list of independent SNPs for PRS construction, we perform linkage disequilibrium (LD)-based SNP clumping for the GWAS results via PLINK. With the default window size (250 kb), we vary the clumping parameter R^2 and set it to 0.05, 0.1, 0.2, 0.3, 0.5, and 0.9. Smaller R^2 results in more stringent selection and more filtered SNPs by clumping. The clumped SNPs are used to construct PRS on testing data.

The results are displayed in Supplementary Figures 12-17, which show similar patterns as in Figure 3. Specifically, GWAS-PRS has consistent performance across the m/p ratio and the clumping parameter R^2 . The prediction accuracy is around $\sqrt{n/(n+p)}$, where p is the number of clumped SNPs given $R^2 = 0.05$ ($p \approx 150,000$). In addition, threshold-PRS may have high prediction accuracy only when the genetic signals are sparse.

5. PRS on Brain Size. We present a real data example to predict the human total brain volume (TBV) with PRS. TBV is generated from brain magnetic resonance imaging (MRI) via advanced normalization tools (Avants et al., 2011) (ANTs) to measure brain size. The UKB sample has 19,629 participants with both TBV and imputed SNP data ($p = 8,944,375$) available after quality controls (Zhao et al., 2019).

We perform a ten-fold analysis where nine folds of the data are used to construct the summary statistics and the remaining one fold of the data is used to measure the PRS prediction accuracy. Twelve PRS are created with p-value thresholds 1, 0.8, 0.5, 0.4, 0.3, 0.2, 0.1, 0.08, 0.05, 0.02, 0.01, and 10^{-3} . Next, we used the GWAS summary statistics estimated from all 19,629 UKB individuals (Zhao et al. (2019), available at <https://med.sites.unc.edu/bigs2/data/gwas-summary-statistics/>) to construct PRS on subjects from three independent studies, including the Human Connectome Project (HCP, $n = 1141$) study, the Pediatric Imaging, Neurocognition, and Genetics (PING, $n = 924$) study, and the Alzheimer’s Disease Neu-

Figure 4: Prediction accuracy (A_P) of PRS for total brain volume in ten-fold UKB data prediction and out-of-UKB prediction.

roimaging Initiative (ADNI, $n = 1248$) study. See supplementary material for more information, such as SNP data quality controls, SNP pruning and demographic information of these studies.

The prediction accuracy is measured by the (partial) R^2 of PRS from the linear regression model for TBV where the effects of age and gender are adjusted for. Figure 4 displays the average out-of-UKB R^2 across HCP, ADNI, and PING, and the R^2 of within-UKB ten-fold prediction. GWAS-PRS constructed with all candidate SNPs have consistent R^2 in within-UKB ten-fold prediction and out-of-UKB prediction, which is around 1.5% (P -value $< 1.92 \times 10^{-6}$). For threshold-PRS, there is a similar pattern that R^2 decreases as the P -value threshold decreases, especially when the cutoff is less than 0.2. Such pattern matches with our asymptotic results of threshold-PRS when SNP signals are dense, clearly suggesting that TBV is a complex polygenic trait. We therefore recommend GWAS-PRS for TBV prediction.

6. Discussion. The PRS is widely used in GWAS for predicting phenotypes and for studying co-heritability of pleiotropy phenotypes (Purcell et al., 2009; Chatterjee, Shi and García-Closas, 2016; Power et al., 2015; Choi, Mak and O'Reilly, 2018; Khera et al., 2018). The purpose of PRS is to aggregate effects from a large number of causal SNPs, each of which has small contribution to the phenotype. However, recent years have witnessed

widespread not so satisfying performances of PRS in real data applications (Bogdan, Baranger and Agrawal, 2018; Zheutlin and Ross, 2018; Márquez-Luna, Loh and Price, 2017; Torkamani, Wineinger and Topol, 2018; Clarke et al., 2016; Mistry et al., 2018a,b; Socrates et al., 2017). Though previous studies (Daetwyler, Villanueva and Woolliams, 2008; Dudbridge, 2013; Chatterjee et al., 2013) have found that the prediction accuracy of PRS is related to (n, p) , the asymptotic properties of PRS is largely unknown. More rigorous investigation done in this paper would greatly help researchers gain better insights on PRS and appreciate the limitations of PRS to avoid over- or under- interpreting their findings. In addition, there exists no theoretical study on how the sparsity of genetic signals affects PRS.

We illustrate for the first time how and why the commonly used marginal screening approaches for polygenic traits may fail in preserving the rank of GWAS signals. The properties of single SNP analysis is closely related to the increasingly recognized spurious correlation problem (Fan, Guo and Hao, 2012; Chen, Fan and Li, 2018; Cai et al., 2011; Cai, Fan and Jiang, 2013; Fan et al., 2018; Fan and Zhou, 2016; Su, 2018) in the statistics community. For polygenic/omnigenic traits, single SNP analysis is always misspecified since the effects of a large number of causal SNPs are not modeled but absorbed into the error term, which can profoundly affect the accuracy of marginal screening, an issue that can be safely ignored for traits with sparse genetic signals.

For GWAS-PRS, our asymptotic results align well with previous studies in genetics community (Daetwyler, Villanueva and Woolliams, 2008; Dudbridge, 2013; Chatterjee et al., 2013), but are more statistically rigorous and general. In addition, we illustrate in Figure 2 that the variation of the PRS performance of A_P increases with the increase of the p/n ratio as well. When the sample size of training data is small, or p/n is large, the prediction accuracy of GWAS-PRS fluctuates around zero with a large variation and the results of PRS should be interpreted with caution. We further generalize our results to threshold-PRS. We highlight how the number of causal SNPs influences the performance of threshold-PRS, and recognize its distinct behaviors under dense and sparse genetic signal scenarios. Threshold-PRS can be powerful for complex traits with sparse genetic signals provided that proper threshold is used. For traits with highly polygenic or omnigenic genetic structures, the general rule is to avoid the use of threshold-PRS but GWAS-PRS. However, the performance of the later is solely determined by the n/p ratio and thus is low when n is not large, leading to a practical paradox. Further, in our theoretical investigation and simulation studies, we assume all causal SNPs are observed, an assumption that likely not holds

for any real GWAS. For real GWAS data with partially genotyped causal SNPs, GWAS-PRS and threshold-PRS are expected to perform even worse.

Last, it is worth mentioning that the limited power of PRS for predicting highly polygenic traits is not unique to PRS, but due to the fundamental analytic challenge in studying such traits. Performed on individual-level data, popular regularization-based methods (such as LASSO (Tibshirani, 1996) or Ridge regression (Hoerl and Kennard, 1970; Tikhonov, 1963)) and the best linear unbiased prediction (BLUP) methods (Chen et al., 2015; Yang et al., 2011) have low prediction power similar to that of PRS unless the genetic signals are sparse where LASSO is clearly advantageous. See simulation results in Supplementary Figures 18-20 with simulation details in the supplementary material.

In conclusion, this study thoroughly investigates the power and limitations of PRS on predicting highly polygenic traits. The research hopefully will increase researchers' awareness on the challenges in studying complex polygenic traits, and the statistics and genetics communities' awareness on the need of developing novel experiments and innovative statistical methods for complex polygenic traits that are inherently difficult to study.

Acknowledgements. We would like to thank Ziliang Zhu, Jingwen Zhang, and Hongtu Zhu for helpful discussions; and Xinlei Mi for independent technical verification. We are deeply grateful to Hongtu Zhu for providing helpful comments and edits; and to Laura Zhou for careful proofreading and feedback. The manuscript has been greatly improved by their generous contributions. This research has been conducted using the UK Biobank resource (application number 22783), subject to a data transfer agreement. We thank Hongtu Zhu, Tengfei Li and other members of the UNC BIG-S2 lab for providing and processing the raw imaging data. We thank the individuals represented in the UK Biobank, PING, HCP, and ADNI studies for their participation and the research teams for their work in collecting, processing and disseminating these datasets for analysis. More information of these studies can be found in the supplementary material.

APPENDIX A: PROOF

In this appendix, we highlight the key steps and important intermediate results to prove the two theorems. More technical details can be found in the supplementary material.

PROPOSITION A1. *Under the polygenic model (2) and Condition (1), if*

$m \rightarrow \infty$ when $n, p \rightarrow \infty$, then we have

$$\frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}}{nm \cdot \sigma_\beta^2} = 1 + o_p(1),$$

$$\frac{Var_2}{\{n^2 m(p-m) + n^2 m(m+n)\} \cdot \sigma_\beta^2} = 1 + o_p(1),$$

where

$$Var_2 = (\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T) \cdot (\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}).$$

Further if $p/n^2 \rightarrow 0$, then we have

$$\frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}}{n^2 m \cdot \sigma_\beta^2} = 1 + o_p(1).$$

By continuous mapping theorem, we have

$$A_P^2 / \left(\frac{n}{n+p} \right) = A_P^2 / \left(\frac{1}{1+\gamma} \right) = 1 + o_p(1).$$

It follows that Theorem 1 is proved for $\alpha \in (0, 2)$. Now consider the case that $p/n^2 \not\rightarrow o(1)$, i.e., $\alpha \in [2, \infty]$. Note that

$$\begin{aligned} & \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)} \\ &= \frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)} - n^2 m \cdot \sigma_\beta^2}{\sqrt{(n^2 m^2 p + 4n^3 m^2) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)}} \\ & \cdot \sqrt{(n^2 m^2 p + 4n^3 m^2) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4) + n^2 m \cdot \sigma_\beta^2} \\ &= \mathcal{O}_p \left[\left\{ (n^2 m^2 p + 4n^3 m^2) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4) \right\}^{1/2} \right] + n^2 m \cdot \sigma_\beta^2. \end{aligned}$$

It follows that

$$\begin{aligned} A_P^2 &= \frac{\mathcal{O}_p \{ (n^2 m^2 p + n^4 m^2) \cdot \sigma_\beta^4 \}}{[nm \cdot \sigma_\beta^2 \cdot \{1 + o(1)\}] \cdot [n^2 m(n+p) \cdot \sigma_\beta^2 \cdot \{1 + o(1)\}]} \\ &= \mathcal{O}_p \left(\frac{n^2 m^2 p + n^4 m^2}{n^3 m^2 p + n^4 m^2} \right) = \mathcal{O}_p \left(\frac{n^2 + c \cdot n^\alpha}{n^2 + c \cdot n^{1+\alpha}} \right) = \mathcal{O}_p \left(\frac{1}{n} \right) \end{aligned}$$

for $\alpha \in [2, \infty]$. Thus, Theorem 1 is proved for $\alpha \in (0, \infty]$.

PROPOSITION A2. *Under the polygenic model (2) and Condition (1), if $m, q_1, q_2 \rightarrow \infty$ when $n, p \rightarrow \infty$, then we have*

$$\frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}}{nm \cdot \sigma_\beta^2} = 1 + o_p(1),$$

$$\frac{Var_2}{\{n^2 m q_2 + n^2 q_1 (m + n)\} \cdot \sigma_\beta^2} = 1 + o_p(1),$$

where

$$Var_2 = (\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T) \cdot$$

$$(\mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)}).$$

Further if $\{m^2(q_1 + q_2)\}/(n^2 q_1^2) \rightarrow 0$, then we have

$$\frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)}}{n^2 q_1 \cdot \sigma_\beta^2} = 1 + o_p(1).$$

By continuous mapping theorem, we have

$$A_P^2 / \left(\frac{n q_1^2}{n m q_1 + q m^2} \right) = 1 + o_p(1).$$

Without loss of generality and for simplicity, we let $\sigma_\beta^2 = 1$, and we note that

$$\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)}$$

$$= \mathcal{O}_p \left[\left\{ n^2 q_1^3 + 2n^2 q_1^2 (m - q_1) + 2n^3 q_1 (m - q_1) + \right. \right.$$

$$\left. \left. n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2 \right\}^{1/2} \right] + n^2 q_1.$$

Then if $\{m^2(q_1 + q_2)\}/(n^2 q_1^2) \not\rightarrow 0$, we have

$$A_P^2 = \mathcal{O}_p \left\{ \frac{m^2(q_1 + q_2) + 2n(m - q_1)q_1 + n^2 q_1^2}{nm^2(q_1 + q_2) + n^2 m q_1} \right\}$$

$$= \mathcal{O}_p \left\{ \frac{m^2(q_1 + q_2) + n^2 q_1^2}{nm^2(q_1 + q_2) + n^2 m q_1} \right\} = \mathcal{O}_p \left(\frac{1}{n} \right) = o_p(1).$$

Thus, Theorem 2 is proved.

REFERENCES

- AVANTS, B. B., TUSTISON, N. J., SONG, G., COOK, P. A., KLEIN, A. and GEE, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* **54** 2033–2044.
- BALDING, D. J. and NICHOLS, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96** 3–12.
- BOGDAN, R., BARANGER, D. A. and AGRAWAL, A. (2018). Polygenic risk scores in clinical psychology: bridging genomic risk to individual differences. *Annual Review of Clinical Psychology* **14** 119–157.
- BOYLE, E. A., LI, Y. I. and PRITCHARD, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169** 1177–1186.
- BYCROFT, C., FREEMAN, C., PETKOVA, D., BAND, G., ELLIOTT, L. T., SHARP, K., MOTYER, A., VUKCEVIC, D., DELANEAU, O., OCONNELL, J. et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562** 203.
- CAI, T., FAN, J. and JIANG, T. (2013). Distributions of angles in random packing on spheres. *The Journal of Machine Learning Research* **14** 1837–1864.
- CAI, T. T., JIANG, T. et al. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *The Annals of Statistics* **39** 1496–1525.
- CHATTERJEE, N., SHI, J. and GARCÍA-CLOSAS, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* **17** 392–406.
- CHATTERJEE, N., WHEELER, B., SAMPSON, J., HARTGE, P., CHANOCK, S. J. and PARK, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature genetics* **45** 400–405.
- CHEN, Z., FAN, J. and LI, R. (2018). Error Variance Estimation in Ultrahigh-Dimensional Additive Models. *Journal of the American Statistical Association* **113** 315–327.
- CHEN, C.-Y., HAN, J., HUNTER, D. J., KRAFT, P. and PRICE, A. L. (2015). Explicit modeling of ancestry improves polygenic risk scores and BLUP prediction. *Genetic Epidemiology* **39** 427–438.
- CHOI, S. W., MAK, T. S. H. and O'REILLY, P. (2018). A guide to performing Polygenic Risk Score analyses. *BioRxiv* 416545.
- CLARKE, T., LUPTON, M., FERNANDEZ-PUJALS, A., STARR, J., DAVIES, G., COX, S., PATTIE, A., LIEWALD, D., HALL, L., MACINTYRE, D. et al. (2016). Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population. *Molecular Psychiatry* **21** 419–425.
- DAETWYLER, H. D., VILLANUEVA, B. and WOOLLIAMS, J. A. (2008). Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* **3** e3395.
- DUDBRIDGE, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genetics* **9** e1003348.
- DUDBRIDGE, F. (2016). Polygenic epidemiology. *Genetic Epidemiology* **40** 268–272.
- FAN, J., GUO, S. and HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 37–65.
- FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 849–911.
- FAN, J. and ZHOU, W.-X. (2016). Guarding against spurious discoveries in high dimen-

- sions. *Journal of Machine Learning Research* **17** 1–34.
- FAN, J., SHAO, Q.-M., ZHOU, W.-X. et al. (2018). Are discoveries spurious? Distributions of maximum spurious correlations and their applications. *The Annals of Statistics* **46** 989–1017.
- FISHER, R. A. (1919). XV.The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* **52** 399–433.
- GE, T., CHEN, C.-Y., NEALE, B. M., SABUNCU, M. R. and SMOLLER, J. W. (2017). Phenome-wide heritability analysis of the UK Biobank. *PLoS Genetics* **13** e1006711.
- GOTTESMAN, I. and SHIELDS, J. (1967). A polygenic theory of schizophrenia. *Proceedings of the National Academy of Sciences* **58** 199–205.
- HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.
- KEMP, J. P., MORRIS, J. A., MEDINA-GOMEZ, C., FORGETTA, V., WARRINGTON, N. M., YOUTEN, S. E., ZHENG, J., GREGSON, C. L., GRUNDBERG, E., TRAJANOSKA, K. et al. (2017). Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nature Genetics* **49** 1468–1475.
- KHERA, A. V., CHAFFIN, M., ARAGAM, K. G., HAAS, M. E., ROSELLI, C., CHOI, S. H., NATARAJAN, P., LANDER, E. S., LUBITZ, S. A., ELLINOR, P. T. et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* **50** 1219–1224.
- LEE, S., WRIGHT, F. A. and ZOU, F. (2011). Control of population stratification by correlation-selected principal components. *Biometrics* **67** 967–974.
- LEE, S. H., DECANDIA, T. R., RIPKE, S., YANG, J., SULLIVAN, P. F., GODDARD, M. E., KELLER, M. C., VISSCHER, P. M., WRAY, N. R., CONSORTIUM, S. P. G.-W. A. S. et al. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics* **44** 247–250.
- MACARTHUR, J., BOWLER, E., CEREZO, M., GIL, L., HALL, P., HASTINGS, E., JUNKINS, H., MCMAHON, A., MILANO, A., MORALES, J. et al. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45** D896–D901.
- MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.
- MÁRQUEZ-LUNA, C., LOH, P.-R. and PRICE, A. L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology* **41** 811–823.
- MISTRY, S., HARRISON, J. R., SMITH, D. J., ESCOTT-PRICE, V. and ZAMMIT, S. (2018a). The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: systematic review. *Schizophrenia Research* **197** 2–8.
- MISTRY, S., HARRISON, J. R., SMITH, D. J., ESCOTT-PRICE, V. and ZAMMIT, S. (2018b). The use of polygenic risk scores to identify phenotypes associated with genetic risk of bipolar disorder and depression: A systematic review. *Journal of Affective Disorders* **234** 148–155.
- PASANIUC, B. and PRICE, A. L. (2017). Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics* **18** 117–127.
- PENROSE, L. (1953). The genetical background of common diseases. *Human Heredity* **4** 257–265.
- POWER, R. A., STEINBERG, S., BJORNSDOTTIR, G., RIETVELD, C. A., ABDELLAOUI, A., NIVARD, M. M., JOHANNESSON, M., GALESLOOT, T. E., HOTTENGA, J. J., WILLEM-

- SEN, G. et al. (2015). Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nature Neuroscience* **18** 953–955.
- PRICE, A. L., PATTERSON, N. J., PLENCE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38** 904–909.
- PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81** 559–575.
- PURCELL, S. M., WRAY, R., STONE, L., VISSCHER, M., O'DONOVAN, C., SULLIVAN, F., SKLAR, P., RUDERFER, M., MCQUILLIN, A., MORRIS, W. et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460** 748–752.
- SHI, H., KICHAIEV, G. and PASANIUC, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics* **99** 139–153.
- SOCRATES, A., BOND, T., KARHUNEN, V., AUVINEN, J., RIETVELD, C., VEIJOLA, J., JARVELIN, M.-R. and O'REILLY, P. (2017). Polygenic risk scores applied to a single cohort reveal pleiotropy among hundreds of human phenotypes. *BioRxiv* 203257.
- SU, W. J. (2018). When is the first spurious variable selected by sequential regression procedures? *Biometrika* **105** 517–527.
- SUDLOW, C., GALLACHER, J., ALLEN, N., BERAL, V., BURTON, P., DANESH, J., DOWNEY, P., ELLIOTT, P., GREEN, J., LANDRAY, M. et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12** e1001779.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288.
- TIKHONOV, A. N. (1963). On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk* **151** 501–504. Russian Academy of Sciences.
- TIMPSON, N. J., GREENWOOD, C. M., SORANZO, N., LAWSON, D. J. and RICHARDS, J. B. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics* **19** 110–124.
- TORKAMANI, A., WINEINGER, N. E. and TOPOL, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* **19** 581–590.
- VILHJÁLMSSON, B. J., YANG, J., FINUCANE, H. K., GUSEV, A., LINDSTRÖM, S., RIPKE, S., GENOVESE, G., LOH, P.-R., BHATIA, G., DO, R. et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics* **97** 576–592.
- VISSCHER, P. M., BROWN, M. A., MCCARTHY, M. I. and YANG, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics* **90** 7–24.
- VISSCHER, P. M., WRAY, N. R., ZHANG, Q., SKLAR, P., MCCARTHY, M. I., BROWN, M. A. and YANG, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101** 5–22.
- WATANABE, K., STRINGER, S., FREI, O., MIRKOV, M. U., POLDERMAN, T. J., VAN DER SLUIS, S., ANDREASSEN, O. A., NEALE, B. M. and POSTHUMA, D. (2018). A global view of pleiotropy and genetic architecture in complex traits. *bioRxiv* 500090.
- WRAY, N. R., YANG, J., HAYES, B. J., PRICE, A. L., GODDARD, M. E. and VISSCHER, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics* **14** 507–515.
- WRAY, N. R., WIJMEGA, C., SULLIVAN, P. F., YANG, J. and VISSCHER, P. M. (2018).

Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* **173** 1573–1580.

YANG, J., BENYAMIN, B., McEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W. et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42** 565–569.

YANG, J., LEE, S. H., GODDARD, M. E. and VISSCHER, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88** 76–82.

YANG, J., BAKSHI, A., ZHU, Z., HEMANI, G., VINKHUYZEN, A. A., LEE, S. H., ROBINSON, M. R., PERRY, J. R., NOLTE, I. M., VAN VLIET-OSTAPTCHOUK, J. V. et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics* **47** 1114–1120.

ZHAO, B., LUO, T., LI, T., LI, Y., ZHANG, J., SHAN, Y., WANG, X., YANG, L., ZHOU, F., ZHU, Z. and ZHU, H. (2019). GWAS of 19,629 individuals identifies novel genetic variants for regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *bioRxiv:586339*.

ZHEUTLIN, A. B. and ROSS, D. A. (2018). Polygenic Risk Scores: What Are They Good For? *Biological Psychiatry* **83** e51–e53.

ZUK, O., HECHTER, E., SUNYAEV, S. R. and LANDER, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109** 1193–1198.

SUPPLEMENTARY MATERIAL

Supplement to: “On PRS for Complex Polygenic Trait Prediction”

(doi: [10.1214/00-AOASXXXXSUPP](https://doi.org/10.1214/00-AOASXXXXSUPP)). We provide additional theoretical details, real data information and simulation results.

DEPARTMENT OF BIostatISTICS
UNIVERSITY OF NORTH CAROLINA
135 DAUER DRIVE
CHAPEL HILL, NORTH CAROLINA 27599
E-MAIL: bingxin@unc.edu
fzou@bios.unc.edu

SUPPLEMENT TO “ON PRS FOR COMPLEX POLYGENIC TRAIT PREDICTION”

1. Intermediate results.

PROPOSITION S1. *Under Condition (1), if $m \rightarrow \infty$ as $n, p \rightarrow \infty$, then we have*

$$\begin{aligned} E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) &= n^2 m \cdot \sigma_\beta^2 \\ \text{Var}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\ &= \{(n^2 m^3 + 4n^3 m^2 + n^2 m^2(p - m)) \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)\} \cdot \{1 + o(1)\} \end{aligned}$$

$$\begin{aligned} E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}) &= nm \cdot \sigma_\beta^2 \\ \text{Var}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}) &= o(n^2 m^2 \cdot \sigma_\beta^4) \end{aligned}$$

$$\begin{aligned} E\{(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T)(\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)})\} \\ = n^2 m(n + m) \cdot \sigma_\beta^2 \cdot \{1 + o(1)\} + n^2 m(p - m) \cdot \sigma_\beta^2 \\ \text{Var}\{(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T)(\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)})\} \\ = o\{n^4 m^2(n + p)^2 \cdot \sigma_\beta^4\}, \end{aligned}$$

where $b_4 < \infty$ is the forth moment of β .

Proposition S1 quantifies the scale of the three terms in A_P . Particularly, for the two variance terms $\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}$ and

$$(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T)(\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}),$$

the expected values can respectively dominate the corresponding standard error for any ratios among (p, m, n) . However, for the covariance term

$$\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)},$$

its standard error may or may be dominated by its expected value depending on p/n . Following Proposition S1, by Markov's inequality, for any constant $k > 0$, we have

$$\Pr\left(\left|\frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}}{nm \cdot \sigma_\beta^2} - 1\right| \geq k\right) \leq \frac{\text{Var}\left(\frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}}{nm \cdot \sigma_\beta^2}\right)}{k^2} = \frac{\text{Var}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)})}{n^2 m^2 \cdot \sigma_\beta^4 k^2} = o(1),$$

and

$$\begin{aligned}
 & \Pr \left\{ \left| \frac{(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T)(\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)})}{n^2 m(n+p) \cdot \sigma_\beta^2} - 1 \right| \geq k \right\} \\
 & \leq \text{Var} \left\{ \frac{(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T)(\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)})}{n^2 m(n+p) \cdot \sigma_\beta^2} \right\} \cdot \frac{1}{k^2} \\
 & = \frac{\text{Var} \{ (\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T)(\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \}}{n^4 m^2 (n+p)^2 \cdot \sigma_\beta^4 k^2} \\
 & = o(1),
 \end{aligned}$$

and

$$\begin{aligned}
 & \Pr \left(\left| \frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}}{n^2 m \cdot \sigma_\beta^2} - 1 \right| \geq k \right) \\
 & \leq \text{Var} \left(\frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}}{n^2 m \cdot \sigma_\beta^2} \right) \cdot \frac{1}{k^2} \\
 & = \frac{\{n^2 m^3 + n^2 m^2 (p-m)\} \cdot \{1 + o(1)\}}{n^4 m^2 \cdot \sigma_\beta^4 k^2} = \frac{p}{n^2 k^2} \cdot \{1 + o(1)\}.
 \end{aligned}$$

It follows that Proposition A1 is proved. More generally, if training and testing data have different sample sizes, denoted as n and n_z , respectively, we have the following results.

PROPOSITION S2. *Under Condition (1), if $m \rightarrow \infty$ as $n, n_z, p \rightarrow \infty$, then we have*

$$\begin{aligned}
 & E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) = nn_z m \cdot \sigma_\beta^2 \\
 & \text{Var}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) = \\
 & \{ (nn_z m^2 p + 2n^2 n_z m^2 + 2nn_z^2 m^2) \cdot \sigma_\beta^4 + n^2 n_z^2 m \cdot (b_4 - \sigma_\beta^4) \} \cdot \{1 + o(1)\}
 \end{aligned}$$

$$\begin{aligned}
 & E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}) = n_z m \cdot \sigma_\beta^2 \\
 & \text{Var}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}) = o(n_z^2 m^2 \cdot \sigma_\beta^4)
 \end{aligned}$$

$$\begin{aligned}
 & E \left\{ (\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T)(\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \right\} \\
 & = nn_z m(n+m) \cdot \sigma_\beta^2 \cdot \{1 + o(1)\} + nn_z m(p-m) \cdot \sigma_\beta^2 \\
 & \text{Var} \left\{ (\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T)(\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \right\} \\
 & = o\{n^2 n_z^2 m^2 (n+p)^2 \cdot \sigma_\beta^4\}.
 \end{aligned}$$

By Markov's inequality and continuous mapping theorem again, we have the following results.

PROPOSITION S3. *Under Condition (1), if $m \rightarrow \infty$ as $n, n_z, p \rightarrow \infty$, then we have*

$$\frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}}{n_z m \cdot \sigma_\beta^2} = 1 + o_p(1)$$

$$\frac{(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T)(\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)})}{[nn_z m(p - m) + nn_z m(n + m)] \cdot \sigma_\beta^2} = 1 + o_p(1).$$

If further $p/(nn_z) \rightarrow 0$, then we have

$$\frac{\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}}{nn_z m \cdot \sigma_\beta^2} = 1 + o_p(1),$$

and it follows that

$$A_P^2 / \left(\frac{n}{n + p} \right) = 1 + o_p(1).$$

When $p/(nn_z) \not\rightarrow 0$, i.e., $\alpha \in [1, \infty]$, we note that

$$\begin{aligned} & \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)} \\ &= \mathcal{O}_p \left\{ \sqrt{nn_z m^2 p + 2n^2 n_z m^2 + 2nn_z^2 m^2 + n^2 n_z^2 m(b_4 - \sigma_\beta^4)} \right\} + nn_z m \cdot \sigma_\beta^2 \\ &= \mathcal{O}_p \left\{ (n^{1/2} n_z^{1/2} m p^{1/2} + nn_z m) \cdot \sigma_\beta^2 \right\}. \end{aligned}$$

It follows that

$$A_P^2 = \frac{\mathcal{O}_p(nn_z m^2 p + n^2 n_z^2 m^2)}{[n_z m \cdot \{1 + o(1)\}] \cdot [nn_z m(n + p) \cdot \{1 + o(1)\}]} = \mathcal{O}_p \left(\frac{p + nn_z}{n_z p + nn_z} \right) = \mathcal{O}_p \left(\frac{1}{n_z} \right).$$

The results of threshold-PRS can also be derived in a similar way. Without loss of generality, we set $\sigma_\beta^2 = 1$ in later steps.

PROPOSITION S4. *Under Conditions (1), if $m, q_1, q_2 \rightarrow \infty$ as $n, p \rightarrow \infty$, then we have*

$$\begin{aligned} E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)}) &= n^2 q_1 \\ \text{Var}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)}) &= \{n^2 q_1^3 + 4n^3 q_1^2 \\ &+ n^4 q_1(b_4 - 1) + 2n^2 q_1^2(m - q_1) + 2n^3 q_1(m - q_1) + n^2(m - q_1)^2 q_1 + n^2 m^2 q_2\} \cdot \{1 + o(1)\} \\ E(\text{Var}_2) &= \{n^2 q_1(n + m) + n^2 q_2 m\} \cdot \{1 + o(1)\} \\ \text{Var}(\text{Var}_2) &= o[\{n^2 q_1(n + m) + n^2 q_2 m\}^2]. \end{aligned}$$

By Markov's inequality and continuous mapping theorem again, Proposition A2 is proved.

2. Technical details. The following technical details are useful to prove our theoretical results. Most of them involve calculating the asymptotic expectation of the trace of the product of multiple large random matrices. We use the definition of matrix trace and apply the combination theory to calculate the total variations. The results provided below may also benefit other research questions involving similar calculations.

2.1. GWAS-PRS. First moment of covariance term

$$\begin{aligned} & E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\ &= \sigma_\beta^2 \cdot E(\text{tr}(\mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)})) \\ &= \sigma_\beta^2 \cdot E\left(\sum_{i=1}^n \sum_{j=1}^n \sum_{k_1=1}^m \sum_{k_2=1}^m Z_{ik_1} X_{jk_1} Z_{ik_2} X_{jk_2}\right) \\ &= \sigma_\beta^2 \cdot E\left(\sum_{i=1}^n \sum_{j=1}^n \sum_{k_1=1}^m Z_{ik_1}^2 X_{jk_1}^2\right) = \sigma_\beta^2 \cdot n^2 m \end{aligned}$$

$$\begin{aligned} & E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\ &= \sigma_\beta^2 \cdot E(\text{tr}(\mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)})) \\ &= \sigma_\beta^2 \cdot E\left(\sum_{i=1}^n \sum_{j=1}^n \sum_{k_2=1}^m \sum_{k_1=1}^m Z_{(2)ik_2} X_{(2)jk_2} X_{(1)jk_1} Z_{(1)ik_1}\right) = 0 \end{aligned}$$

Thus, we have $E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z} \mathbf{X}^T \mathbf{X}_{(1)} \beta_{(1)}) = \sigma_\beta^2 \cdot n^2 m$.

First moment of variance term I

$$\begin{aligned} & E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}) \\ &= E\{E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)} | Z)\} = E(\text{tr}(\mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \cdot \mathbf{I}_m \cdot \sigma_\beta^2) + 0) \\ &= \sigma_\beta^2 \cdot E(\text{tr}(\mathbf{Z}_{(1)} \mathbf{Z}_{(1)}^T)) = \sigma_\beta^2 \cdot E\left(\sum_{i=1}^n \sum_{j=1}^m Z_{ij}^2\right) = \sigma_\beta^2 \cdot nm \end{aligned}$$

First moment of variance term II

$$\begin{aligned}
& E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\
&= \sigma_\beta^2 \cdot E(\text{tr}(\mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)})) \\
&= \sigma_\beta^2 \cdot E\left(\sum_{i=1}^n \sum_{j=1}^m \sum_{k_1=1}^m \sum_{l_1=1}^n \sum_{k_2=1}^m \sum_{l_2=1}^n Z_{ik_1} Z_{ik_2} X_{l_1 k_1} X_{l_1 j} X_{l_2 k_2} X_{l_2 j}\right) \\
&= \sigma_\beta^2 \cdot E\left(\sum_{i=1}^n \sum_{k_1=k_2=j}^m \sum_{l_1 \neq l_2}^{n(n-1)} Z_{ik_1}^2 X_{l_1 k_1}^2 X_{l_2 k_1}^2 + \sum_{i=1}^n \sum_{k_1=k_2 \neq j}^{m(m-1)} \sum_{l_1=l_2}^n Z_{ik_1}^2 X_{l_1 k_1}^2 X_{l_1 j}^2\right. \\
&\quad \left.+ \sum_{i=1}^n \sum_{k_1=k_2=j}^m \sum_{l_1=l_2}^n Z_{ik_1}^2 X_{l_1 k_1}^4\right) \\
&= \sigma_\beta^2 \cdot n^2 m(n+m+c_4-2) = \sigma_\beta^2 \cdot n^2 m(n+m) \cdot \{1+o(1)\},
\end{aligned}$$

where $c_4 = E(X_{11}^4) < \infty$.

$$\begin{aligned}
& E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\
&= \sigma_\beta^2 \cdot E(\text{tr}(\mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)})) \\
&= \sigma_\beta^2 \cdot E\left(\sum_{c=1}^n \sum_{i=1}^m \sum_{k=1}^{p-m} \sum_{l=1}^n \sum_{q=1}^{p-m} \sum_{r=1}^n Z_{(2)ck} Z_{(2)cq} X_{(2)lk} X_{(2)rq} X_{(1)li} X_{(1)ri}\right) \\
&= \sigma_\beta^2 \cdot n^2 m(p-m)
\end{aligned}$$

$$E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) = 0$$

Thus, we have

$$E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X} \mathbf{Z}^T \mathbf{Z} \mathbf{X}^T \mathbf{X}_{(1)} \beta_{(1)}) = \sigma_\beta^2 \cdot n^2 m(n+m) \cdot \{1+o(1)\} + \sigma_\beta^2 \cdot n^2 m(p-m)$$

Second moment of covariance term

$$\begin{aligned}
& E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\
&= E(\text{tr}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)})) \\
&= E\left(\sum_{c=1}^n \sum_{d=1}^n \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m \sum_{l=1}^n \sum_{q=1}^m \sum_{r=1}^m \sum_{s=1}^n \sum_{t=1}^m Z_{ck} Z_{dq} Z_{dr} Z_{ct} X_{lk} X_{li} X_{sr} X_{sj} \beta_i \beta_t \beta_j\right) \\
&= \{(n^4 m^2 + 4n^3 m^2 + n^2 m^3) \cdot \sigma_\beta^4 + mn^4 \cdot (b_4 - \sigma_\beta^4)\} \cdot \{1+o(1)\}
\end{aligned}$$

$$\begin{aligned}
& E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\
&= E(\text{tr}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)})) \\
&= E\left(\sum_{c=1}^n \sum_{d=1}^n \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^{p-m} \sum_{l=1}^n \sum_{q=1}^m \sum_{r=1}^{p-m} \sum_{s=1}^n \sum_{t=1}^m \right. \\
&\quad \left. Z_{(2)ck} Z_{(2)dr} Z_{(1)dq} Z_{(1)ct} X_{(2)lk} X_{(2)sr} X_{(1)li} X_{(1)sj} \beta_q \beta_i \beta_t \beta_j\right) \\
&= n^2(p-m)\{m^2 + (\sigma_\beta^4 - 1) \cdot m\} \cdot \sigma_\beta^4 = n^2 m^2 (p-m) \cdot \sigma_\beta^4 \cdot \{1 + o(1)\}
\end{aligned}$$

$$E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)}) \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) = 0$$

It follows that

$$\begin{aligned}
& \text{Var}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\
&= [\{n^2 m^3 + 4n^3 m^2 + n^2 m^2 (p-m)\} \cdot \sigma_\beta^4 + n^4 m \cdot (b_4 - \sigma_\beta^4)] \cdot \{1 + o(1)\}.
\end{aligned}$$

Second moment of variance term I

$$\begin{aligned}
& E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}) \\
&= E(\text{tr}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)})) \\
&= E\left(\sum_{c=1}^n \sum_{d=1}^n \sum_{i=1}^m \sum_{j=1}^m \sum_{k=1}^m \sum_{l=1}^m X_{ci} X_{dj} X_{dk} X_{cl} \beta_i \beta_j \beta_k \beta_l\right) \\
&= E\left\{\sum_{c=d}^n \left(\sum_{i=j \neq k=l}^{m(m-1)} X_{ci}^2 X_{ck}^2 \beta_i^2 \beta_k^2 + \sum_{i=k \neq j=l}^{m(m-1)} X_{ci}^2 X_{cj}^2 \beta_i^2 \beta_j^2 + \sum_{i=l \neq j=k}^{m(m-1)} X_{ci}^2 X_{cj}^2 \beta_i^2 \beta_j^2\right.\right. \\
&\quad \left.\left.+ \sum_{i=l=j=k}^m X_{ci}^4 \beta_i^4\right) + \sum_{c \neq d}^{n(n-1)} \left(\sum_{i=l \neq k=j}^{m(m-1)} X_{ci}^2 X_{dj}^2 \beta_i^2 \beta_j^2 + \sum_{i=l=k=j}^m X_{ci}^2 X_{dj}^2 \beta_i^4\right)\right\} \\
&= \sigma_\beta^4 \cdot n^2 m^2 + nm \cdot \{2m\sigma_\beta^4 + n(b_4 - \sigma_\beta^4) + c_4 b_4 - 2\sigma_\beta^4 - b_4\},
\end{aligned}$$

where $b_4 = E(\beta_1^4) < \infty$. It follows that

$$\text{Var}(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \beta_{(1)}) = nm \cdot \{2m\sigma_\beta^4 + n(b_4 - \sigma_\beta^4) + c_4 b_4 - 2\sigma_\beta^4 - b_4\} = o(n^2 m^2 \cdot \sigma_\beta^4).$$

Second moment of variance term II

$$\begin{aligned}
& E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)}) \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\
&= E(\text{tr}(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)})) \\
&= E\left(\sum_{c=1}^n \sum_{d=1}^n \sum_{h=1}^m \sum_{k=1}^m \sum_{l=1}^n \sum_{i=1}^m \sum_{q=1}^m \sum_{r=1}^n \sum_{s=1}^m \sum_{t=1}^m \sum_{u=1}^n \sum_{w=1}^m \sum_{a=1}^m \sum_{b=1}^n \right. \\
&\quad \left. Z_{ck} Z_{dq} Z_{dt} Z_{ca} X_{lk} X_{lh} X_{rq} X_{ri} X_{ut} X_{us} X_{ba} X_{bw} \beta_h \beta_i \beta_s \beta_w\right)
\end{aligned}$$

To have non-zero means, we need possible combinations with the following patterns:

- either $Z_{..}^2 Z_{..}^2$ or $Z_{..}^4$
- one of $X_{..}^2 X_{..}^2 X_{..}^2 X_{..}^2$, $X_{..}^2 X_{..}^2 X_{..}^4$, $X_{..}^4 X_{..}^4$, $X_{..}^6 X_{..}^2$, and $X_{..}^8$
- either $\beta^2 \beta^2$ or β^4 .

After tedious calculations, we have

$$\begin{aligned} & E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\ &= \sigma_\beta^4 \cdot \{n^4 m^2 (n+m)^2\} \cdot \{1 + o(1)\}. \end{aligned}$$

Next, we have

$$\begin{aligned} & E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\ &= E(\text{tr}(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)})) \\ &= E\left(\sum_{c=1}^n \sum_{d=1}^n \sum_{h=1}^m \sum_{k=1}^{p-m} \sum_{l=1}^n \sum_{i=1}^m \sum_{q=1}^{p-m} \sum_{r=1}^n \sum_{s=1}^m \sum_{t=1}^{p-m} \sum_{u=1}^n \sum_{w=1}^m \sum_{a=1}^{p-m} \sum_{b=1}^n \right. \\ &\quad \left. Z_{(2)ck} Z_{(2)dq} Z_{(2)dt} Z_{(2)ca} X_{(2)lk} X_{(2)rq} X_{(2)ut} X_{(2)ba} X_{(1)lh} X_{(1)ri} X_{(1)us} X_{(1)bw} \beta_h \beta_i \beta_s \beta_w\right) \\ &= \sigma_\beta^4 \cdot \{n^4 m^2 (p-m)^2\} \cdot \{1 + o(1)\} \end{aligned}$$

and

$$\begin{aligned} & E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\ &= E(\text{tr}(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)})) \\ &= E\left(\sum_{c=1}^n \sum_{d=1}^n \sum_{h=1}^m \sum_{k=1}^{p-m} \sum_{l=1}^n \sum_{i=1}^m \sum_{q=1}^m \sum_{r=1}^n \sum_{s=1}^{p-m} \sum_{t=1}^n \sum_{u=1}^m \sum_{w=1}^m \sum_{a=1}^m \sum_{b=1}^n \right. \\ &\quad \left. Z_{(2)ck} Z_{(1)dq} Z_{(2)dt} Z_{(1)ca} X_{(2)lk} X_{(2)ut} X_{(1)lh} X_{(1)rq} X_{(1)ri} X_{(1)us} X_{(1)ba} X_{(1)bw} \beta_h \beta_i \beta_s \beta_w\right) \\ &= \sigma_\beta^4 \cdot n(p-m) \{ (m^3 - 3m^2 + 2m)(n^2 + 2n) + (m^2 - m)(n^3 - n^2 + 2(c_4 - 1)n + 4n^2 + 4(c_4 - 1)n + b_4 n^2) + b_4 m n^3 \} \\ &= \sigma_\beta^4 \cdot \{n^3 m^2 (p-m)(m+n)\} \cdot \{1 + o(1)\}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} & E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\ &= o\{\sigma_\beta^4 \cdot n^4 m^2 (n+p)^2\} \end{aligned}$$

and

$$\begin{aligned} & E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\ &= E(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)}) = 0. \end{aligned}$$

It follows that

$$\begin{aligned} & \text{Var}\left\{(\beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \mathbf{Z}_{(1)}^T + \beta_{(1)}^T \mathbf{X}_{(1)}^T \mathbf{X}_{(2)} \mathbf{Z}_{(2)}^T)(\mathbf{Z}_{(1)} \mathbf{X}_{(1)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(2)} \mathbf{X}_{(2)}^T \mathbf{X}_{(1)} \beta_{(1)})\right\} \\ &= o\{\sigma_\beta^4 \cdot n^4 m^2 (n+p)^2\}. \end{aligned}$$

The results of different sample sizes can be similarly derived and are ignored. Without loss of generality, we set $\sigma_\beta^2 = 1$ for simplicity below for threshold-PRS.

2.2. Threshold-PRS. First moment of covariance term

$$\begin{aligned} C_1 &= \beta_{(1)}^T \mathbf{Z}_{(1)}^T (\mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(1)} \beta_{(1)} + \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\ &= \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(1)} \beta_{(1)} + \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)} \\ &= \beta_{(1)}^T \mathbf{Z}_{(1)}^T (\mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)} + \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)} + \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)}) \\ &= \beta_{(11)}^T \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)} + \beta_{(11)}^T \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)} \\ &+ \beta_{(12)}^T \mathbf{Z}_{(12)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)} + \beta_{(12)}^T \mathbf{Z}_{(12)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)} \\ &+ \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)} = C_{11} + C_{12} + C_{13} + C_{14} + C_{15} \end{aligned}$$

$$\text{Thus, } E(C_1) = E(C_{11}) = E(\beta_{(11)}^T \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)}) = n^2 q_1$$

First moment of variance term II

$\text{Var}_2 =$

$$\begin{aligned}
 & (\beta_{(11)}^T \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T + \beta_{(12)}^T \mathbf{X}_{(12)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T + \beta_{(11)}^T \mathbf{X}_{(11)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T + \beta_{(12)}^T \mathbf{X}_{(12)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T) \cdot \\
 & (\mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)} + \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)} + \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(11)} \beta_{(11)} + \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(12)} \beta_{(12)}) \\
 & = \beta_{(11)}^T \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)} + \beta_{(11)}^T \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)} \\
 & + \beta_{(11)}^T \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(11)} \beta_{(11)} + \beta_{(11)}^T \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(12)} \beta_{(12)} \\
 & + \beta_{(12)}^T \mathbf{X}_{(12)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)} + \beta_{(12)}^T \mathbf{X}_{(12)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)} \\
 & + \beta_{(12)}^T \mathbf{X}_{(12)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(11)} \beta_{(11)} + \beta_{(12)}^T \mathbf{X}_{(12)}^T \mathbf{X}_{(11)} \mathbf{Z}_{(11)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(12)} \beta_{(12)} \\
 & + \beta_{(11)}^T \mathbf{X}_{(11)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)} + \beta_{(11)}^T \mathbf{X}_{(11)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)} \\
 & + \beta_{(11)}^T \mathbf{X}_{(11)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(11)} \beta_{(11)} + \beta_{(11)}^T \mathbf{X}_{(11)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(12)} \beta_{(12)} \\
 & + \beta_{(12)}^T \mathbf{X}_{(12)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)} + \beta_{(12)}^T \mathbf{X}_{(12)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)} \\
 & + \beta_{(12)}^T \mathbf{X}_{(12)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(11)} \beta_{(11)} + \beta_{(12)}^T \mathbf{X}_{(12)}^T \mathbf{X}_{(21)} \mathbf{Z}_{(21)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(12)} \beta_{(12)} \\
 & = a^2 + ab + ac + ad + ba + b^2 + bc + bd + ca + cb + c^2 + cd + da + db + dc + d^2
 \end{aligned}$$

$\text{E}(\text{Var}_2)$

$$\begin{aligned}
 & = \text{E}(a^2 + ab + ac + ad + ba + b^2 + bc + bd + ca + cb + c^2 + cd + da + db + dc + d^2) \\
 & = \text{E}(a^2 + b^2 + c^2 + d^2) \\
 & = \{n^2 q_1 (n + q_1)\} \cdot \{1 + o(1)\} + n^2 (m - q_1) q_1 + n^2 q_1 q_2 + n^2 (m - q_1) q_2 \\
 & = \{n^2 q_1 (n + m + q_2) + n^2 q_2 (m - q_1)\} \cdot \{1 + o(1)\} \\
 & = \{n^2 q_1 (n + m) + n^2 q_2 m\} \cdot \{1 + o(1)\}
 \end{aligned}$$

Second moment of covariance term

$$\begin{aligned}
 \text{E}(C_1 C_1) & = \text{E}\{(C_{11} + C_{12} + C_{13} + C_{14} + C_{15})(C_{11} + C_{12} + C_{13} + C_{14} + C_{15})\} \\
 & = \text{E}(C_{11}^2 + C_{12}^2 + C_{13}^2 + C_{14}^2 + C_{15}^2)
 \end{aligned}$$

$$\begin{aligned}
 \text{E}(C_{11}^2) & = \text{E}(\beta_{(11)}^T \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)} \beta_{(11)}^T \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)}) \\
 & = n^4 q_1^2 + \{(n^2 q_1^3 + 4n^3 q_1^2) + n^4 q_1 (b_4 - 1)\} \cdot \{1 + o(1)\}
 \end{aligned}$$

$$\begin{aligned}
 E(C_{12}^2) &= E(\beta_{(11)}^T \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)} \beta_{(11)}^T \mathbf{Z}_{(11)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)}) \\
 &= E\left(\sum_{c=1}^n \sum_{d=1}^n \sum_{i=1}^{m-q_1} \sum_{j=1}^{m-q_1} \sum_{k=1}^{q_1} \sum_{l=1}^n \sum_{q=1}^{q_1} \sum_{r=1}^{q_1} \sum_{s=1}^n \sum_{t=1}^{q_1} \right. \\
 &\quad \left. Z_{(11)ck} Z_{(11)dq} Z_{(11)dr} Z_{(11)ct} X_{(11)lk} X_{(11)sr} X_{(12)li} X_{(12)sj} \beta_{(11)q} \beta_{(11)t} \beta_{(12)i} \beta_{(12)j} \right) \\
 &= \{n^2 q_1^2 (m - q_1) + n^3 q_1 (m - q_1)\} \cdot \{1 + o(1)\}
 \end{aligned}$$

Similarly, we have

$$\begin{aligned}
 E(C_{13}^2) &= E(\beta_{(12)}^T \mathbf{Z}_{(12)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)} \beta_{(12)}^T \mathbf{Z}_{(12)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(11)} \beta_{(11)}) \\
 &= \{n^2 q_1^2 (m - q_1) + n^3 q_1 (m - q_1)\} \cdot \{1 + o(1)\} = E(C_{12}^2)
 \end{aligned}$$

$$\begin{aligned}
 E(C_{14}^2) &= E(\beta_{(12)}^T \mathbf{Z}_{(12)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)} \beta_{(12)}^T \mathbf{Z}_{(12)}^T \mathbf{Z}_{(11)} \mathbf{X}_{(11)}^T \mathbf{X}_{(12)} \beta_{(12)}) \\
 &= n^2 (m - q_1)^2 q_1
 \end{aligned}$$

$$E(C_{15}^2) = E(\beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)} \beta_{(1)}^T \mathbf{Z}_{(1)}^T \mathbf{Z}_{(21)} \mathbf{X}_{(21)}^T \mathbf{X}_{(1)} \beta_{(1)}) = n^2 m^2 q_2.$$

Thus, we have

$$\begin{aligned}
 E(C_1 C_1) &= E(C_{11}^2 + C_{12}^2 + C_{13}^2 + C_{14}^2 + C_{15}^2) \\
 &= n^4 q_1^2 + \{(n^2 q_1^3 + 4n^3 q_1^2) + n^4 q_1 (b_4 - 1)\} \cdot \{1 + o(1)\} \\
 &\quad + 2 \cdot \{n^2 q_1^2 (m - q_1) + n^3 q_1 (m - q_1)\} \cdot \{1 + o(1)\} + n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2.
 \end{aligned}$$

It follows that

$$\begin{aligned}
 \text{Var}(C_1) &= \{n^2 q_1^3 + 4n^3 q_1^2 + n^4 q_1 (b_4 - 1) + 2n^2 q_1^2 (m - q_1) + \\
 &\quad 2n^3 q_1 (m - q_1) + n^2 (m - q_1)^2 q_1 + n^2 m^2 q_2\} \cdot \{1 + o(1)\}.
 \end{aligned}$$

Second moment of variance term II

$$\begin{aligned}
 E(\text{Var}_2 \text{Var}_2) &= E\{(a^2 + ab + ac + ad + ba + b^2 + bc + bd + ca + cb + c^2 + cd + da + db + dc + d^2)^2\} \\
 &= E\{a^4 + b^4 + c^4 + d^4 + 2a^2 b^2 + 2a^2 c^2 + 2a^2 d^2 + 2b^2 c^2 + 2b^2 d^2 + 2c^2 d^2\} \\
 &= E(a^4) + E(b^4) + E(c^4) + E(d^4) + E(2a^2 b^2 + 2a^2 c^2 + 2a^2 d^2 + 2b^2 c^2 + 2b^2 d^2 + 2c^2 d^2) \\
 &= \{n^2 q_1 (n + q_1)\}^2 + o(n^4 q_1^4 + n^5 q_1^3 + n^6 q_1^2) + \{n^2 (m - q_1) q_1\}^2 \cdot \{1 + o(1)\} \\
 &\quad + (n^2 q_1 q_2)^2 \cdot \{1 + o(1)\} + \{n^2 (m - q_1) q_2\}^2 \cdot \{1 + o(1)\} + o\{(n^2 q_1 (n + m) + n^2 q_2 m)^2\}
 \end{aligned}$$

It follows that $\text{Var}(\text{Var}_2) = o[\{n^2 q_1 (n + m) + n^2 q_2 m\}^2]$.

3. Simulation setups: BLUP, Ridge, and LASSO. We perform simple simulations to numerically illustrate the prediction accuracy of BLUP, ridge and LASSO given different m/p ratios.

As in Case 1, we simulate $p = 100,000$ uncorrelated SNPs from $N(0, 1)$. We vary the number of causal SNPs m and set it to 100, 1000, 10,000 and 50,000. The causal SNP effects $\beta_{(1)} \sim MVN(0, \mathbf{I}_m)$. The linear polygenic model (2) is used to generate phenotype $\mathbf{y}$. The sample size is set to 1000 or 10,000 for training data, and 1000 for testing data. A total of 100 replications are conducted for each simulation condition. The results of BLUP prediction are shown in Supplementary Figure 18. The performance of BLUP is consistent across different m/p ratios, and is slightly larger than $\sqrt{n/(n+p)}$.

For LASSO and ridge regression, we use smaller (n, p, m) to reduce the computational burden. We simulate $p = 10,000$ uncorrelated SNPs from $N(0, 1)$. We vary the number of causal SNPs m and set it to 100, 1000, 25,000 and 5,000. The sample size is set to 2000 for training data and 1000 for testing data. Other settings are exactly the same as those in BLUP. LASSO has nearly perfect performance when the signals are sparse. However, its performance is consistently lower than $\sqrt{n/(n+p)}$ when the signals are dense (Supplementary Figure 19). The pattern of ridge regression is very similar to that of PRS. When all the SNPs are used, ridge regression has consistent performance regardless of the m/p ratios, which is slightly larger than $\sqrt{n/(n+p)}$ (Supplementary Figure 20).

4. Real data analysis. The detailed cohort information, data quality control, and GWAS procedures can be found in Zhao et al. (2019), Supplementary Table 1 gives a brief summary of the demographic information of five datasets. In each testing dataset, we use PLINK toolset (Purcell et al., 2007) to generate risk scores in testing data by summarizing across pruned (window size 50, step 5, $R^2 = 0.2$) SNP alleles, weighed by their effect sizes estimated from training data. The UKB effect size estimates are taken from Zhao et al. (2019) and have been made publicly available at <https://med.sites.unc.edu/bigs2/data/gwas-summary-statistics/>. The association between each pair of polygenic profile and ROI volume is estimated and tested in linear regression, adjusting for the age and gender. The additional variance of ROI volume that can be explained by PRS is used to measure the prediction power.

Data Acknowledgement. Part of data collection and sharing for this project was funded by the Alzheimers Disease Neuroimaging initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (De-

SUPPLEMENTARY TABLE 1
Demographic information of five datasets.

Dataset	Sample size	Mean age (s.d.)	Age range	Proportion of male
UK Biobank	19,629	62.51 (7.47)	(40 80)	0.473
ADNI	1248	74.18 (7.16)	(55 92)	0.566
HCP	1141	28.82 (3.68)	(22 36)	0.542
PING	924	12.28 (4.99)	(3 21)	0.518
PNC	1492	21.14 (3.71)	(14 29)	0.515

partment of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering and through generous contributions from the following: Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Part of the data collection and sharing for this project was funded by the Pediatric Imaging, Neurocognition and Genetics Study (PING) (U.S. National Institutes of Health Grant RC2DA029475). PING is funded by the National Institute on Drug Abuse and the Eunice Kennedy Shriver National Institute of Child Health & Human Development. PING data are disseminated by the PING Coordinating Center at the Center for Human Development, University of California, San Diego. HCP data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

PING Methods. Part of the data used in the preparation of this article were obtained from the Pediatric Imaging, Neurocognition and Genetics (PING) Study database (<http://ping.chd.ucsd.edu/>). PING was launched in 2009 by the National Institute on Drug Abuse (NIDA) and the Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD) as a 2-year project of the American Recovery and Reinvestment Act. The primary goal of PING has been to create a data resource of highly standardized and carefully curated magnetic resonance imaging (MRI) data, comprehensive genotyping data, and developmental and neuropsychological assessments for a large cohort of developing children aged 3 to 20 years. The scientific aim of the project is, by openly sharing these data, to amplify the power and productivity of investigations of healthy and disordered development in children, and to increase understanding of the origins of variation in neurobehavioral phenotypes. For up-to-date information, see <http://ping.chd.ucsd.edu/>.

ADNI Methods. Data used in the preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a 60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1,

ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

Pediatric Imaging, Neurocognition and Genetics (PING) Authors. Connor McCabe¹, Linda Chang², Natacha Akshoomoff³, Erik Newman¹, Thomas Ernst², Peter Van Zijl⁴, Joshua Kuperman⁵, Sarah Murray⁶, Cinnamon Bloss⁶, Mark Appelbaum¹, Anthony Gamst¹, Wesley Thompson³, Hauke Bartsch⁵.

Alzheimer's Disease Neuroimaging Initiative (ADNI) Authors. Michael Weiner⁷, Paul Aisen¹, Ronald Petersen⁸, Clifford R. Jack Jr⁸, William Jagust⁹, John Q. Trojanowski¹⁰, Arthur W. Toga¹¹, Laurel Beckett¹², Robert C. Green¹³, Andrew J. Saykin¹⁴, John Morris¹⁵, Leslie M. Shaw¹⁰, Zaven Khachaturian¹⁶, Greg Sorensen¹⁷, Maria Carrillo¹⁸, Lew Kuller¹⁹, Marc Raichle¹⁵, Steven Paul²⁰, Peter Davies²¹, Howard Fillit²², Franz Hefti²³, Davie Holtzman¹⁵, M. Marcel Mesulman²⁴, William Potter²⁵, Peter J. Snyder²⁶, Adam Schwartz²⁷, Tom Montine²⁸, Ronald G. Thomas¹, Michael Donohue¹, Sarah Walter¹, Devon Gessert¹, Tamie Sather¹, Gus Jiminez¹, Danielle Harvey¹², Matthew Bernstein⁸, Nick Fox²⁹, Paul Thompson¹¹, Norbert Schuff⁷, Charles DeCarli¹², Bret Borowski⁸, Jeff Gunter⁸, Matt Senjem⁸, Prashanthi Vemuri⁸, David Jones⁸, Kejal Kantarci⁸, Chad Ward⁸, Robert A. Koeppe³⁰, Norm Foster³¹, Eric M. Reiman³², Kewei Chen³², Chet Mathis¹⁹, Susan Landau⁹, Nigel J. Cairns¹⁵, Erin Householder¹⁵, Lisa Taylor-Reinwald¹⁵, Virginia M.Y. Lee¹⁰, Magdalena Korecka¹⁰, Michal Figurski¹⁰, Karen Crawford¹¹, Scott Neu¹¹, Tatiana M. Foroud¹⁴, Steven Potkin³³, Li Shen¹⁴, Kelley Faber¹⁴, Sungeun Kim¹⁴, Kwangsik Nho¹⁴, Leon Thal¹, Richard Frank³⁴, Neil Buckholtz³⁵, Marilyn Albert³⁶, John Hsiao³⁵.

¹UC San Diego, La Jolla, CA 92093, USA. ²U Hawaii, Honolulu, HI 96822, USA. ³Department of Psychiatry, University of California, San Diego, La Jolla, California 92093, USA. ⁴Kennedy Krieger Institute, Baltimore, MD 21205, USA. ⁵Multimodal Imaging Laboratory, Department of Radiology, University of California San Diego, La Jolla, California 92037, USA. ⁶Scripps Translational Science Institute, La Jolla, CA 92037, USA. ⁷UC San Francisco, San Francisco, CA 94143, USA. ⁸Mayo Clinic, Rochester, MN 55905, USA. ⁹UC Berkeley, Berkeley, CA 94720-5800, USA. ¹⁰U Pennsylvania, Philadelphia, PA 19104, USA. ¹¹USC, University of Southern California, Los Angeles, CA 90033, USA. ¹²UC Davis, Davis, CA

95616, USA. ¹³Brigham and Women's Hospital/Harvard Medical School, Boston MA 02115, USA. ¹⁴Indiana University, Indianapolis, IN 46202-5143, USA. ¹⁵Washington University St. Louis, St. Louis, MO 63130, USA. ¹⁶Prevent Alzheimers Disease 2020, Rockville, MD 20850, USA. ¹⁷Siemens ¹⁸Alzheimers Association, Chicago, IL 60601, USA. ¹⁹University of Pittsburgh, Pittsburgh, PA 15260, USA. ²⁰Cornell University, Ithaca, NY 14850, USA. ²¹Albert Einstein College of Medicine of Yeshiva University, Bronx, NY 10461, USA. ²²AD Drug Discovery Foundation, New York, NY 10019, USA. ²³Acumen Pharmaceuticals, Livermore, California 94551, USA. ²⁴Northwestern University, Evanston, IL 60208, USA. ²⁵National Institute of Mental Health, Bethesda, MD 20892-9663, USA. ²⁶Brown University, Providence, RI 02912, USA. ²⁷Eli Lilly, Indianapolis, Indiana 46285, USA. ²⁸University of Washington, Seattle, WA 98195, USA. ²⁹University of London, London WC1E 7HU, UK. ³⁰University of Michigan, Ann Arbor, MI 48109, USA. ³¹University of Utah, Salt Lake City, UT 84112, USA. ³²Banner Alzheimers Institute, Phoenix, AZ 85006, USA. ³³UC Irvine, Irvine, CA 92697, USA. ³⁴General Electric ³⁵National Institute on Aging/National Institutes of Health, Bethesda, MD 20892, USA. ³⁶The Johns Hopkins University, Baltimore, MD 21218, USA.

References.

- PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81** 559–575.
- ZHAO, B., LUO, T., LI, T., LI, Y., ZHANG, J., SHAN, Y., WANG, X., YANG, L., ZHOU, F., ZHU, Z. and ZHU, H. (2019). GWAS of 19,629 individuals identifies novel genetic variants for regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *bioRxiv:586339*.

Supplementary Figure 1: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. We set $p=100,000$ and $n=1000$ in both training and testing data.

POLYGENIC RISK SCORE PREDICTION

17

Supplementary Figure 2: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. SNP data are independently sampled from $\{0, 1, 2\}$. We set $p=100,000$ and $n=1000$ in both training and testing data.

Supplementary Figure 3: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios. SNP data are independently sampled from $\{0, 1, 2\}$. We set $p=100,000$ and $n=10,000$ in training data and $n=1000$ in testing data.

Supplementary Figure 4: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the m causal SNPs are known and are only considered as candidates in constructing PRS. SNP data are independently sampled from $\{0, 1, 2\}$. We set $p=100,000$ and $n=1000$ in both training and testing data.

Supplementary Figure 5: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the m causal SNPs are known and are only considered as candidates in constructing PRS. SNP data are independently sampled from $\{0, 1, 2\}$. We set $p=100,000$ and $n=10,000$ in training data and $n = 1000$ in testing data.

Supplementary Figure 6: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the heritability $h^2=0.5$. SNP data are independently sampled from $\{0, 1, 2\}$. We set $p=100,000$ and $n=1000$ in both training and testing data.

Supplementary Figure 7: Prediction accuracy (A_P) of threshold-PRS across different m/p when the heritability $h^2=0.5$. SNP data are independently sampled from $\{0, 1, 2\}$. We set $p=100,000$ and $n=10,000$ in training data and $n=1000$ in testing data.

Supplementary Figure 8: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when population substructure exists in the SNP data. SNP data are independently sampled from $\{0, 1, 2\}$. We set $p=100,000$ and $n=1000$ in both training and testing data.

Supplementary Figure 9: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when population substructure exists in the SNP data. SNP data are independently sampled from $\{0, 1, 2\}$. We set $p=100,000$ and $n=10,000$ in training data and $n=1000$ in testing data.

Supplementary Figure 10: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the effects of causal SNPs are not i.i.d. Normal. SNP data are independently sampled from $\{0, 1, 2\}$. We set $p=100,000$ and $n=1000$ in both training and testing data.

Supplementary Figure 11: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios when the effects of causal SNPs are not i.i.d. Normal. SNP data are independently sampled from $\{0, 1, 2\}$. We set $p=100,000$ and $n=10,000$ in training data and $n=1000$ in testing data.

Supplementary Figure 12: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios in the UKB data simulation. There are $p = 461,488$ SNPs and $n = 10,000$ individuals in the training data. Clumping R^2 parameter is set to be 0.05. $p \approx 150,000$ SNPs pass LD-based clumping and are candidates SNPs to construct PRS on $n = 1000$ subjects in testing data.

Supplementary Figure 13: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios in the UKB data simulation. There are $p = 461,488$ SNPs and $n = 10,000$ individuals in the training data. Clumping R^2 parameter is set to be 0.1. $p \approx 190,000$ SNPs pass LD-based clumping and are candidates SNPs to construct PRS on $n = 1000$ subjects in testing data.

Supplementary Figure 14: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios in the UKB data simulation. There are $p = 461,488$ SNPs and $n = 10,000$ individuals in the training data. Clumping R^2 parameter is set to be 0.2. $p \approx 250,000$ SNPs pass LD-based clumping and are candidates SNPs to construct PRS on $n = 1000$ subjects in testing data.

Supplementary Figure 15: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios in the UKB data simulation. There are $p = 461,488$ SNPs and $n = 10,000$ individuals in the training data. Clumping R^2 parameter is set to be 0.3. $p \approx 290,000$ SNPs pass LD-based clumping and are candidates SNPs to construct PRS on $n = 1000$ subjects in testing data.

Supplementary Figure 16: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios in the UKB data simulation. There are $p = 461,488$ SNPs and $n = 10,000$ individuals in the training data. Clumping R^2 parameter is set to be 0.5. $p \approx 360,000$ SNPs pass LD-based clumping and are candidates SNPs to construct PRS on $n = 1000$ subjects in testing data.

Supplementary Figure 17: Prediction accuracy (A_P) of threshold-PRS across different m/p ratios in the UKB data simulation. There are $p = 461,488$ SNPs and $n = 10,000$ individuals in the training data. Clumping R^2 parameter is set to be 0.9. $p \approx 430,000$ SNPs pass LD-based clumping and are candidates SNPs to construct PRS on $n = 1000$ subjects in testing data.

Supplementary Figure 18: Prediction accuracy (A_P) of BLUP across different m/p ratios. We set $p=100,000$ and m to be 100, 1000, 10,000, and 50,000, respectively. In (a), we set $n=1000$ in both training and testing data, and in (b), we set $n=10,000$ in training data and $n=1000$ in testing data.

Supplementary Figure 19: Prediction accuracy (A_P) of LASSO across different m/p ratios. We set $p=10,000$ and $n=2000$ in training data and $n=1000$ in testing data.

Supplementary Figure 20: Prediction accuracy (A_P) of ridge regression across different m/p ratios. We set $p=10,000$ and $n=2000$ in training data and $n=1000$ in testing data.