Journal of Genetics (2022)101:14
https://doi.org/10.1007/s12041-021-01349-4

RESEARCH ARTICLE

Competing analytical strategies of combining associated SNPs for estimating genetic risks

ARUNABHA MAJUMDAR[1,2] and SAURABH GHOSH[2]*

[1]*Department of Mathematics, Indian Institute of Technology Hyderabad, Kandi 502 284, India*
[2]*Human Genetics Unit, Indian Statistical Institute, Kolkata 700 108, India*
*For correspondence. E-mail: saurabh@isical.ac.in.

Received 6 September 2020; revised 15 August 2021; accepted 4 October 2021

Abstract. In genomewide association study (GWAS) of a complex phenotype, a large number of variants, many with small effect sizes, are found to contribute to the variability of the phenotype. Subsequent to the identification of such variants in a GWAS, it is of interest to estimate the risk jointly conferred by the variants. We propose three different strategies of combining the risk SNPs to calculate an allele dosage score. Using simulations, we evaluate the different measures of allele dosage score with respect to the risk prediction accuracy of a binary trait and the proportion of variance explained for a quantitative trait. For a binary trait, an allele dosage score defined based on log odds ratio performs marginally better than the other two measures. For a quantitative trait, the measure based on the standardized slope coefficient in linear regression of the trait on single-nucleotide polymorphism (SNP) genotypes performs better than the measures using the weights proportional to log *P*-value and the proportion of variance explained. We demonstrate the utility of these measures using a real data on type 2 diabetes and fasting blood sugar level in a south Indian population.

Keywords. genomewide association study; allele dosage score; regression; variance explained; prediction.

Introduction

Genomewide association scans are designed to identify single-nucleotide polymorphisms (SNPs) that are in linkage disequilibrium with loci modulating the phenotype under study. Since complex phenotypes are controlled by multiple loci, primarily having minor effect sizes, it is of interest to assess the risk jointly conferred by variants that exhibit significant evidence of marginal association. To integrate the allelic effects from multiple associated SNPs, a simple strategy is to calculate the mean number of risk alleles combining all the associated SNPs together. However, since the effect sizes are in general expected to vary across the SNPs, each SNP should be differentially weighted in the combined score. Hence, it is more appropriate to consider a weighted mean of the number of risk alleles at the associated SNPs, the weights being proportional to an appropriate measure of the effect sizes of the associated SNPs (Chauhan *et al.* 2010). We define an 'allele dosage score' (ADS) as a suitable weighted mean of the number of risk alleles across the associated SNPs which is also often termed as a polygenic risk score (PRS) in the literature (IS Consortium 2009; Dudbridge 2013; Peyrot *et al.* 2014; Power *et al.* 2015; Mega *et al.* 2015; Kendler 2016; Hachiya *et al.* 2017; Torkamani *et al.* 2018; Khera *et al.* 2018; Martin *et al.* 2019; Duncan *et al.* 2019). Such a score can be viewed as the effective number of risk alleles combining all the associated SNPs.

In the recent years, the utility of ADS has been explored from the perspective of both association testing and risk prediction of a disease. If ADS derived by combining a set of SNPs is found to be associated with a complex trait, it indicates a joint contribution of the set of markers in modulating the trait. IS Consortium (2009) successfully applied allele dosage scores to test for an association with schizophrenia at the genomewide level. Another promising aspect of allele dosage scores lies in its use for predicting the risk of a disease or individual quantitative trait values (Wray *et al.* 2007). Dudbridge (2013) demonstrated that, besides significant association findings of allele dosage scores with different phenotypes, negative association results can be explained by inadequate sample sizes and a moderate increase in the sample size would lead to more fruitful association analyses.

Published online: 25 January 2022

The weight of each risk SNP included in ADS represents a measure of the SNP's effect size. For a case–control phenotype, the effect size is, in general, measured by log odds ratio (OR) estimated from logistic regression of the phenotype on SNP genotypes. Alternatively, one can measure the effect size by -log(*P*-value) obtained from testing the genetic association between the phenotype and SNP using logistic/linear regression for a binary/quantitative phenotype. Similarly, another possibility is to consider the weight to be proportional to the proportion of the variance of the phenotype explained due to the SNP genotypes. The *P*-value and the proportion of variance can be estimated while testing for association between the SNP and an arbitrary type of phenotype (e.g., an ordinal phenotype). However, it may be challenging to define an OR for nonbinary phenotypes. For a case–control phenotype, even though the *P*-value and the proportion of variance explained depend on the estimated log OR, the relationship is often nonlinear. Hence, it is interesting to study if a specific choice of the weights used in ADS performs better for a specific model of the disease/phenotype. Thus, it is important to compare the performance of ADS defined based on such different possible measures of the effect size. We consider three such choices of weights used in allele dosage score, and explore the relative performance of the measures with respect to risk prediction of a binary trait and trait value prediction for a quantitative trait.

For an association study of a binary trait, the standard choice is to consider weights proportional to the log-transformed OR (log(OR)) of the associated SNPs which are estimated from a logistic regression. In the context of a quantitative trait, we need to replace OR with an appropriate alternative measure. For a quantitative trait, we can construct a measure of ADS which is analogous to using log(OR) for a binary trait by assigning weights proportional to the standardized estimate of the slope regression coefficient corresponding to an associated SNP in a linear regression of the trait on the SNP genotypes. For a quantitative trait as well as a binary trait, we propose two new definitions of ADS, one based on weights proportional to the log-transformed *P*-values, and the other based on the proportion of variance of the trait explained by the associated SNPs. We attempt to find the best measure of ADS in different scenarios for a binary as well as a quantitative trait.

Using extensive simulations based on a cross validation approach, we compare the performances of different measures of allele dosage scores including the standard measure based on log(OR) in predicting the risk of a binary trait based on the commonly used logistic regression framework, and the three measures discussed above in explaining the proportion of variance of a quantitative trait. The logistic regression is a popular framework for predicting the risk of a disease outcome based on explanatory variables (Prentice and Pyke 1979). It is known that logistic models fitted based on samples selected via disease ascertainment can produce biased estimates of individual-level risks at the population level (Prentice and Pyke 1979). Hence, it may not be appropriate, though often commonly practiced, to estimate individual-level genetic risks from the same genomewide case–control study with prefixed sample sizes as the one used to identify SNPs associated with the disease, since such study designs are usually biased in terms of oversampling of cases compared to the general population, and hence overestimating the risks. We elucidate this phenomenon by an explicit example that genetic risks should be estimated based on random sampling from the population (i.e., without any ascertainment based on the affection status of an individual). Finally, we explore the performance of our proposed measures of ADS using a real data on type 2 diabetes mellitus (T2DM) and fasting blood sugar (FBS) levels in a south Indian population (Ramya *et al.* 2013).

Materials and methods

Data description and measures of allele dosage scores

For a binary trait, we consider genotype data on n_1 cases and n_2 controls $(n_1 + n_2 = n, \text{total sample size})$, while for a quantitative trait, we consider genotype and phenotype data on n individuals. We assume that, k SNPs were found to be significantly associated at the genomewide level with the phenotype under consideration based on a genomewide scan. Suppose that the alleles at the jth SNP, $j = 1, 2, \ldots, k$, are A_j and a_j, such that A_j is the allele that increases the risk of a disease in case of a binary trait, or induces higher (lower) values of a quantitative trait depending on the direction of risk of a disease associated with the quantitative trait. For an individual, suppose r_j denotes the number of risk alleles (assuming values 0, 1, or 2) at the jth associated SNP, $j = 1, 2, \ldots, k$.

Let $w_1, w_2, \ldots, w_k$, subject to the condition $\sum_{j=1}^{k} w_j = 1$, represent the relative effect sizes of the k associated SNPs. For an individual, define the ADS reflecting the combined risk conferred by the k associated SNPs, as $\sum_{j=1}^{k} w_j r_j$. Thus, the score can vary between 0 and 2.

Suppose, the test for association between a SNP and binary trait is based on OR. Let the estimated OR of the k associated SNPs be denoted by $OR_1, OR_2, \ldots, OR_k$, the *P*-values corresponding to these tests be given by $PV_1, PV_2, \ldots, PV_k$, and the estimated proportions of the total variance of the phenotype explained by the associated SNPs be denoted by $VE_1, VE_2, \ldots, VE_k$, respectively. We denote the allele dosage score defined based on log(OR), log(*P* value), and the estimated proportion of variance explained by ADS_{OR}, ADS_{PV} and ADS_{VE}, respectively. The allele dosage scores of an individual

possessing $r_1, r_2, \ldots, r_k$ risk alleles at k associated SNPs are defined as follows:

$$ADS_{OR} = \frac{\sum_{j=1}^{k} \log\,(OR_j) r_j}{\sum_{j=1}^{k} \log\,(OR_j)},$$

$$ADS_{PV} = \frac{\sum_{j=1}^{k} \log\,(PV_j) r_j}{\sum_{j=1}^{k} \log\,(PV_j)},$$

$$ADS_{VE} = \frac{\sum_{j=1}^{k} VE_j r_j}{\sum_{j=1}^{k} VE_j}.$$

By default, $w_1, \ldots, w_k$ used in ADS_{PV} and ADS_{VE} are always positive. For ADS_{OR}, we encode the genotypes of each SNP with respect to the specific allele such that the OR estimated based on the training sample is always positive. This ensures that $w_1, \ldots, w_k$ used in ADS_{OR} are always positive and sum up to one. We note that the denominator used in the above definitions are not, in general, used in the common definition of PRS. Our definitions provide an additional interpretation of ADS as a weighted number of risk alleles obtained by combining the risk SNPs (taking a value between 0 and 2).

For a quantitative trait, the linear regression of phenotype on SNP genotype provides a test for association. Suppose, $E(Y|G) = \alpha + \beta G$, where Y and G denote the values of the quantitative trait and genotype variable, respectively. Without the loss of generality, the slope coefficient β in the linear regression can always be considered to be positive (by encoding the genotype variable appropriately). At the jth SNP, let the estimate of the standardized β coefficient be denoted by β_j, $j = 1, 2, \ldots, k$. Then, the allele dosage score based on $\beta_1, \beta_2, \ldots, \beta_k$, can be defined as:

$$ADS_{Beta} = \frac{\sum_{j=1}^{k} \beta_j r_j}{\sum_{j=1}^{k} \beta_j}.$$

It is clear that the P-values corresponding to the tests of association can be obtained from the estimated standardized slope coefficients in the above linear regression. For a quantitative trait, the other two measures of allele dosage score, ADS_{PV} and ADS_{VE}, can be defined to be same as those for a binary trait discussed earlier. Here we note that the log-transformed P-value and the proportion of variance explained are not a linear function of the estimated slope regression coefficient. Hence, it is important to evaluate the comparative performance of these different definitions of the allele dosage scores.

We can analytically show that ADS_{Beta} of a quantitative trait is analogous to ADS_{OR} defined for a binary trait. Based on the logistic regression framework, let the risk of a disease conditioned on minor allele at a SNP be modelled as:

$$P(\text{case}|X = x) = \frac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}} = 1 - \frac{1}{1 + \exp\{\alpha + \beta x\}},$$

where, $X = 1$, if an allele is a minor allele, and $X = 0$, otherwise. It follows from the above model that, $\text{OR} = \frac{P(\text{ case } |X=1)}{P(\text{ case } |X=0)} \times \frac{P(\text{ control } |X=0)}{P(\text{ control } |X=1)} = \exp\,(\beta)$ and hence, $\log(\text{OR}) = \beta$. We note that the parameter β in the logistic model signifies the change in the risk of the disease when X changes from 0 to 1. Similarly, for a quantitative trait, the slope coefficient β in the linear regression represents the differential change in the value of the trait for a unit increase in the genotype coding value. Thus, both log(OR) for a binary trait and the slope coefficient β in the linear regression of a quantitative trait are similar with respect to quantifying the effect size of a SNP.

Next, we outline the procedures for estimating different measure of effect sizes of a binary and quantitative trait. Let Y denote the disease status and G denote the genotype of a SNP of interest. We consider the logistic regression: $P(Y = 1) = \frac{\exp(\alpha+\beta G)}{1+\exp(\alpha+\beta G)}$. Based on the training sample, the log OR β is estimated. The P-value of testing $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ is computed from the logistic regression. The proportion of variance of Y explained due to the SNP genotypes (VE) is estimated as the ratio: $\frac{U}{V}$. Here, U is the variance of the estimated risk $[\frac{\exp(\hat{\alpha}+\hat{\beta}G)}{1+\exp(\hat{\alpha}+\hat{\beta}G)}]$ of the individuals in the sample, and $V = \hat{p}(1 - \hat{p})$ is the sample variance of Y, where $\hat{p}$ is the estimated disease prevalence in the population.

Similarly, let Y denote a quantitative trait. We use a linear regression model: $Y = \alpha + \beta G + e$, where e is the random error component. The standardized slope coefficient is estimated as $\frac{\hat{\beta}}{s.e.(\hat{\beta})}$, where $\hat{\beta}$ is the maximum likelihood estimates (m.l.e.) of β. The P-value of testing $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ is obtained. The proportion of variance of Y explained is estimated as: $\frac{\text{var}(Y)-\text{var}(\hat{e})}{\text{var}(Y)}$, where $\text{var}(\hat{e})$ is the variance of the residuals estimated from the linear regression.

Simulation design

To evaluate the performances of different measures of allele dosage score with respect to the accuracy of predicting the risk of a disease, we consider, throughout our simulation study, a case–control design comprising 10,000 individuals. The popular case–control study design usually involves a prefixed number of cases and controls. However, this design may not be appropriate for estimating the vector of penetrance parameters. Since the disease prevalence in a

population is in general much lower than 50%, cases are substantially over-sampled in a study design consisting of equal number of cases and controls which results in overestimation of the penetrance vector. We provide an explicit example to demonstrate this in the 'Appendix' section. Hence, we do not prefix the number of cases (or controls) and assign the affection status to the individuals at random according to the overall disease prevalence in the population.

We describe the simulation strategy and the criteria to compare between the three measures of allele dosage score. We adopt a cross-validation approach, and hence, divide the set of all individuals in the sample at random into a training sample and a test sample. In our simulation study, the training sample comprises 75% of the individuals in the sample and the test sample is composed of the remaining 25% of the individuals. The weights involved in the different measures of allele dosage score are estimated based on the training sample, and are used for calculating the allele dosage scores of the individuals in both test and training samples. We create 1000 sets of training and test samples at random from the simulated data.

For a binary trait, the regression equation to predict the risk based on allele dosage scores is obtained from the training sample. The disease risks of the individuals in the test sample are predicted from this equation based on the allele dosage scores. We evaluate the efficiency of each measure of allele dosage score by the mean area under the receiver's operating characteristic (ROC) curve (AUC), where the mean was computed across 1000 test samples.

For a quantitative trait, we consider a total sample of 4000 individuals. We consider 75% of the individuals as the training sample and the remaining 25% individuals as the test sample. The weights used in different measures of allele dosage scores are estimated based on the training sample. We estimate the linear regression equation to predict the trait value based on the allele dosage score from the training sample, and use it to predict the trait values of individuals in test sample using their allele dosage scores. To evaluate the efficiency of a measure of allele dosage score, we estimate the proportion of total variance of the trait values for individuals in the test sample explained by their allele dosage scores. Thus, we estimate the proportion of variance explained as $\frac{V(\hat{Y})}{V(Y)}$. Here, $\hat{Y} = \hat{a} + \hat{b}X$, where X denotes the allele dosage score for an individual in the test sample, and the coefficients $(\hat{a}, \hat{b})$ in the prediction equation were estimated based on the training sample. Finally, we calculate the mean of the estimated proportions of variance explained for 1000 test samples. Further details about the simulation strategies and models are provided below.

We consider a disease and define a variable $Y = 1$, if an individual is affected, and $Y = 0$, otherwise. We consider two unlinked disease loci with possible alleles (D_1, d_1) and (D_2, d_2), respectively. We define the penetrances in terms of the nine genotype combinations at the two disease risk SNPs. Suppose the penetrance vector $(f_1, f_2, \ldots, f_9)$ is defined as: $f_1 = P(\text{case}|D_1D_1, D_2D_2)$, $f_2 = P(\text{case}|D_1D_1, D_2d_2), f_3 = P(\text{case}|D_1D_1, d_2d_2), \ldots, f_9 = P(\text{case}|d_1d_1, d_2d_2)$. Thus, for an individual with genotypes D_1D_1 and D_2d_2 at the risk loci, the theoretical risk of having the disease is f_2. Given a choice of minor allele frequency for a risk SNP, we simulate the genotypes at the SNP assuming Hardy–Weinberg equilibrium (HWE). For each individual, we simulate the case–control status according to the penetrance probability conditioned on the combination of genotypes at the risk SNPs.

We test for association at the risk SNPs using the OR test under the framework of logistic regression based on the training sample and hence, the estimates of the log OR, *P*-values and the proportion of the variance explained are obtained as described above.

For a binary trait, we also considered another simulation model comprising five risk SNPs. We consider a logistic model of the disease status. We simulate the log OR for the five risk SNPs at random from Uniform(1.1, 1.5). Thus, we consider $P(Y = 1) = \frac{\exp(\alpha + \sum_{j=1}^{5} \beta_j G_j)}{1 + \exp(\alpha + \sum_{j=1}^{5} \beta_j G_j)}$. We simulate the minor allele frequency from Uniform(0.05, 0.5) for the five SNPs. For each risk SNP, we simulate the genotype data assuming HWE. Next, we assign the case–control status to individuals in the sample according to the probability of being affected as determined by the logistic model given the genotypes at risk SNPs for the individual.

While estimating the proportion of variance explained by a risk SNP, we estimate $E(Y|X = x) = P(Y = 1|X = x) = \frac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}}$, where X is the number of minor alleles at the marker locus. The estimates of α and β are plugged in the expression of $E(Y|X = x)$. The sample variance of estimates of $E(Y|X = x)$ for individuals in the training sample is used as an estimate of the variance of binary trait in training sample explained by the risk SNP. The logistic regression is used to predict the risk of the disease for an individual based on the allele dosage score of the individual. So, given the value of the allele dosage score $U = u$, the risk of the disease is estimated by: $P(Y = 1|U = u) = \frac{\exp\{a + bu\}}{1 + \exp\{a + bu\}}$. We estimate a and b based on the training sample.

Next, we discuss the simulation strategy for a quantitative trait. Suppose for n unrelated individuals, we have the data of a quantitative trait Y. We consider two unlinked quantitative trait loci (QTLs). Suppose, the two possible alleles are A_1, a_1, at the 1st QTL, and are A_2, a_2, at the 2nd QTL, with $P(A_1) = p_1$ and $P(A_2) = p_2$. We also define, $q_1 = 1 - p_1$, $q_2 = 1 - p_2$. Let us denote the genotypes at the 1st QTL by: $G_1 = A_1A_1, G_2 = A_1a_1, G_3 = a_1a_1$, and the genotypes at the 2nd QTL by: $H_1 = A_2A_2, H_2 = A_2a_2, H_3 = a_2a_2$.

The conditional means of Y conditioning on the joint genotypes at the two QTLs are given by: $E(Y|G_1H_1) =$

Table 1. Mean AUC and percentage of times of best AUC obtained by a measure of allele dosage score for disease models 1, 2, 3, 4, 5 and 6.

	ADS_{OR}		ADS_{PV}		ADS_{VE}	
Disease models	Mean AUC (SD)	Per cent of times highest AUC	Mean AUC (SD)	Per cent of times highest AUC	Mean AUC (SD)	Per cent of times highest AUC
1	0.61 (0.02)	97.1%	0.61 (0.02)	2.5%	0.61 (0.02)	0.4%
2	0.65 (0.02)	98.8%	0.64 (0.02)	1.2%	0.64 (0.02)	0.0%
3	0.65 (0.02)	99.0%	0.65 (0.02)	1.0%	0.65 (0.02)	0.0%
4	0.64 (0.02)	79.9%	0.63 (0.02)	19.4%	0.63 (0.02)	0.7%
5	0.65 (0.02)	92.1%	0.65 (0.02)	7.8%	0.65 (0.02)	0.1%
6	0.65 (0.02)	79.9%	0.64 (0.02)	19.4%	0.64 (0.02)	0.7%

The choice of the penetrance vector $(f_1, f_2, \ldots, f_9)$ is (0.30, 0.25, 0.20, 0.25, 0.16, 0.10, 0.20, 0.10, 0.05). The allele frequencies at the risk loci are given by (0.1, 0.1), (0.2, 0.2), (0.3, 0.3), (0.2, 0.1), (0.3, 0.2), (0.3, 0.1), for disease models 1, 2, 3, 4, 5, 6, respectively. The penetrance vector is defined as: $f_1 = P(\text{case}|D_1D_1, D_2D_2), f_2 = P(\text{case}|D_1D_1, D_2d_2), f_3 = P(\text{case}|D_1D_1, d_2d_2), \ldots, f_9 = P(\text{case}|d_1d_1, d_2d_2)$.

Table 2. Mean AUC and percentage of times of best AUC obtained by a measure of allele dosage score for disease models 7, 8, 9, 10, 11 and 12.

	ADS_{OR}		ADS_{PV}		ADS_{VE}	
Disease models	Mean AUC (SD)	Per cent of times highest AUC	Mean AUC (SD)	Per cent of times highest AUC	Mean AUC (SD)	Per cent of times highest AUC
7	0.63 (0.02)	75.7%	0.63 (0.02)	21.3%	0.63 (0.02)	3.0%
8	0.65 (0.02)	80.3%	0.65 (0.02)	18.6%	0.65 (0.02)	1.1%
9	0.65 (0.02)	77.4%	0.65 (0.02)	22.3%	0.65 (0.02)	0.3%
10	0.65 (0.02)	67.5%	0.65 (0.02)	32.2%	0.65 (0.02)	0.3%
11	0.66 (0.02)	79.9%	0.66 (0.02)	20.1%	0.66 (0.02)	0.0%
12	0.66 (0.02)	80.3%	0.66 (0.02)	19.7%	0.66 (0.02)	0.0%

The choice of the penetrance vector $(f_1, f_2, \ldots, f_9)$ is (0.30, 0.27, 0.25, 0.20, 0.16, 0.12, 0.15, 0.10, 0.05). The allele frequencies at the risk loci are given by (0.1, 0.1), (0.2, 0.2), (0.3, 0.3), (0.2, 0.1), (0.3, 0.2), (0.3, 0.1), for disease models 7, 8, 9, 10, 11, 12, respectively. The penetrance vector is defined as: $f_1 = P(\text{case}|D_1D_1, D_2D_2), f_2 = P(\text{case}|D_1D_1, D_2d_2), f_3 = P(\text{case}|D_1D_1, d_2d_2), \ldots, f_9 = P(\text{case}|d_1d_1, d_2d_2)$.

$\alpha_1 + \alpha_2$, $E(Y|G_1H_2) = \alpha_1 + \beta_2, E(Y|G_1H_3) = \alpha_1 - \alpha_2$; $E(Y|G_2H_1) = \beta_1 + \alpha_2, E(Y|G_2H_2) = \beta_1 + \beta_2, E(Y|G_2H_3) = \beta_1 - \alpha_2$; $E(Y|G_3H_1) = -\alpha_1 + \alpha_2$, $E(Y|G_3H_2) = -\alpha_1 + \beta_2, E(Y|G_3H_3) = -\alpha_1 - \alpha_2$. So, the conditional means of Y can be expressed in a general form: $E(Y|G_iH_j) = \mu_{ij} = a_i + b_j$, where $a_1 = \alpha_1, a_2 = \beta_1, a_3 = -\alpha_1$, and $b_1 = \alpha_2, b_2 = \beta_2, b_3 = -\alpha_2$.

We note that, with respect to the marginal effect of a QTL (e.g., a_1, a_2, a_3 for first QTL), the QTL has a codominant effect when $a_2 = (a_1 + a_3)/2$. We considered choices of a_1, a_2, a_3 such that both scenarios can be evaluated: $a_2 = (a_1 + a_3)/2$ and $a_2 \neq (a_1 + a_3)/2$; for example, $a_1 = 1.0, a_2 = 0, a_3 = -1.0$, and $a_1 = 1.0, a_2 = 0.5, a_3 = -1.0$.

Hence, the means of Y conditioned only on the genotypes at the first QTL can be derived as: $E(Y|G_1) = \alpha_1 + \{\alpha_2(p_2{}^2 - q_2{}^2) + 2\beta_2 p_2 q_2\}$, $E(Y|G_2) = \beta_1 + \{\alpha_2(p_2{}^2 - q_2{}^2) + 2\beta_2 p_2 q_2\}$, $E(Y|G_3) = -\alpha_1 + \{\alpha_2(p_2{}^2 - q_2{}^2) + 2\beta_2 p_2 q_2\}$. Similarly, the means of Y conditioned only on the genotypes at the second QTL can be obtained.

In the above model of QTL effects on the quantitative trait, the marginal QTL effects were added to obtain the joint QTL effect on the trait. To explore models which deviate from such additivity, we also considered the following epistatic model: $E(Y|G_iH_j) = \mu_{ij} = a_i + b_j + \frac{1}{2}a_ib_j$.

For an individual, we first simulate the genotypes at the QTLs. To simulate Y, we consider the linear model: $Y = \mu + e$, where $\mu = \mu_{ij}$, according as the joint genotypes at the two QTLs are G_i, H_j, respectively; and e is the random error component with mean 0, and variance σ^2. In our simulations, the distributions of e are considered to be normal and chi-squared with degree of freedom one.

The tests for association are performed using a linear regression of the trait on the genotype of the QTLs. The P-values and the proportions of total variance of Y explained by the QTLs are estimated based on the training sample. We

Table 3. Mean AUC and percentage of times of best AUC obtained by a measure of allele dosage score for disease models 13, 14, 15, 16, 17 and 18.

	ADS_{OR}		ADS_{PV}		ADS_{VE}	
Disease models	Mean AUC (SD)	Per cent of times highest AUC	Mean AUC (SD)	Per cent of times highest AUC	Mean AUC (SD)	Per cent of times highest AUC
13	0.57 (0.02)	95.4%	0.57 (0.02)	3.8%	0.57 (0.02)	0.8%
14	0.6 (0.02)	98.3%	0.6 (0.02)	1.7%	0.6 (0.02)	0.0%
15	0.62 (0.01)	99.0%	0.62 (0.01)	1.0%	0.62 (0.01)	0.0%
16	0.59 (0.02)	75.5%	0.59 (0.02)	23.5%	0.59 (0.02)	1.0%
17	0.61 (0.02)	87.7%	0.61 (0.02)	12.2%	0.61 (0.02)	0.1%
18	0.6 (0.02)	71.4%	0.6 (0.02)	28.0%	0.6 (0.01)	0.6%

The choice of the penetrance vector $(f_1, f_2, \ldots, f_9)$ is (0.35, 0.30, 0.25, 0.30, 0.22, 0.15, 0.25, 0.15, 0.10). The allele frequencies at the risk loci are given by (0.1, 0.1), (0.2, 0.2), (0.3, 0.3), (0.2, 0.1), (0.3, 0.2), (0.3, 0.1), for disease models 13, 14, 15, 16, 17, 18, , respectively. The penetrance vector is defined as: $f_1 = P(\text{case}|D_1D_1, D_2D_2), f_2 = P(\text{case}|D_1D_1, D_2d_2), f_3 = P(\text{case}|D_1D_1, d_2d_2), \ldots, f_9 = P(\text{case}|d_1d_1, d_2d_2)$.

Table 4. Description of different disease models in the simulations considered in tables 1, 2 and 3.

Disease model	Penetrance vector $(f_1, f_2, \ldots, f_9)$	Disease loci allele frequency
1	(0.30, 0.25, 0.20, 0.25, 0.16, 0.10, 0.20, 0.10, 0.05)	(0.1,0.1)
2	Same as model 1	(0.2,0.2)
3	Same as model 1	(0.3,0.3)
4	Same as model 1	(0.2,0.1)
5	Same as model 1	(0.3,0.2)
6	Same as model 1	(0.3,0.1)
7	(0.30, 0.27, 0.25, 0.20, 0.16, 0.12, 0.15, 0.10, 0.05)	(0.1,0.1)
8	Same as model 7	(0.2,0.2)
9	Same as model 7	(0.3,0.3)
10	Same as model 7	(0.2,0.1)
11	Same as model 7	(0.3,0.2)
12	Same as model 7	(0.3,0.1)
13	(0.35, 0.30, 0.25, 0.30, 0.22, 0.15, 0.25, 0.15, 0.10)	(0.1,0.1)
14	Same as model 13	(0.2,0.2)
15	Same as model 13	(0.3,0.3)
16	Same as model 13	(0.2,0.1)
17	Same as model 13	(0.3,0.2)
18	Same as model 13	(0.3,0.1)

The penetrance vector is defined as: $f_1 = P(\text{case}|D_1D_1, D_2D_2), f_2 = P(\text{case}|D_1D_1, D_2d_2), f_3 = P(\text{case}|D_1D_1, d_2d_2), \ldots, f_9 = P(\text{case}|d_1d_1, d_2d_2)$.

also estimate the slope coefficients normalized by their standard errors from the linear regressions at the QTLs based on the training sample. The linear regression equation of Y regressing on the allele dosage score is fitted based on the individuals in the training sample.

Results

Simulation results

For a binary trait, we first consider three different choices of penetrance vector defined in terms of two disease loci jointly, and for each such choice, three possible values of minor allele frequency are considered at each of the risk SNPs: 0.1, 0.2, 0.3. We outline the different disease models defined with respect to different combinations of the penetrance vector and allele frequency at risk SNPs. We carry out the simulations for disease models 1–6 corresponding to the first choice of the penetrance vector, models 7–12 corresponding to the second choice, and models 13–18 corresponding to the third choice.

For a given choice of penetrance, we vary the effect sizes of the disease loci by considering different choices of the allele frequencies at the disease loci. In the first and third choices of the penetrance vector, both the disease loci are assigned equal risk in the sense that, if $f_1, f_2, \ldots, f_9$, are arranged into a (3×3) matrix with its consecutive rows as (f_1, f_2, f_3), (f_4, f_5, f_6), and (f_7, f_8, f_9), the matrix turns out to be a symmetric matrix. The second choice induces

Table 5. Summary of AUC obtained by different measures of allele dosage score for a binary trait simulated with five risk SNPs under a logistic disease model.

	log OR		log *P*-value		Variance explained	
Random seed	Mean AUC (SD)	Per cent of times highest AUC	Mean AUC (SD)	Per cent of times highest AUC	Mean AUC (SD)	Per cent of times highest AUC
1	0.59 (0.01)	84.0%	0.59 (0.01)	10.6%	0.59 (0.01)	5.4%
2	0.6 (0.01)	84.6%	0.6 (0.01)	9.7%	0.6 (0.01)	5.7%
3	0.59 (0.01)	85.1%	0.59 (0.01)	8.4%	0.59 (0.01)	6.5%
4	0.59 (0.01)	80.1%	0.59 (0.01)	12.8%	0.59 (0.01)	7.1%
5	0.59 (0.02)	79.4%	0.59 (0.02)	11.9%	0.59 (0.02)	8.7%
6	0.61 (0.01)	84.1%	0.61 (0.01)	10.8%	0.61 (0.01)	5.1%
7	0.59 (0.01)	78.9%	0.58 (0.01)	12.6%	0.58 (0.01)	8.5%
8	0.6 (0.01)	81.6%	0.59 (0.01)	12.0%	0.59 (0.01)	6.4%
9	0.57 (0.01)	77.1%	0.56 (0.01)	13.9%	0.56 (0.01)	9.0%
10	0.59 (0.01)	82.5%	0.58 (0.01)	10.3%	0.58 (0.01)	7.2%
11	0.56 (0.01)	68.3%	0.56 (0.01)	17.1%	0.56 (0.01)	14.6%
12	0.6 (0.02)	79.9%	0.59 (0.02)	12.4%	0.59 (0.02)	7.7%
13	0.6 (0.01)	85.7%	0.6 (0.01)	9.6%	0.6 (0.01)	4.7%
14	0.62 (0.02)	85.2%	0.61 (0.02)	10.5%	0.61 (0.02)	4.3%
15	0.61 (0.02)	85.7%	0.6 (0.02)	9.2%	0.6 (0.02)	5.1%
16	0.6 (0.01)	86.7%	0.6 (0.01)	9.6%	0.59 (0.01)	3.7%
17	0.61 (0.02)	84.6%	0.6 (0.02)	10.1%	0.6 (0.02)	5.3%
18	0.61 (0.02)	85.2%	0.6 (0.02)	9.0%	0.6 (0.02)	5.8%
19	0.58 (0.01)	79.5%	0.58 (0.01)	13.5%	0.58 (0.01)	7.0%
20	0.62 (0.02)	88.0%	0.61 (0.02)	8.4%	0.61 (0.02)	3.6%

Table 6. Mean variance explained and the percentage of times maximum variance explained of the trait by each measure of allele dosage score for quantitative trait models 1, 2, 3, 4, 5 and 6.

	ADS_{Beta}		ADS_{PV}		ADS_{VE}	
Quantitative trait models	Mean VE	Per cent of times maximum VE	Mean VE	Per cent of times maximum VE	Mean VE	Per cent of times maximum VE
1	15.4% (1.7%)	63.6	15.3% (1.7%)	28.9	15.3% (1.7%)	7.5
2	42% (3%)	65.7	41.9% (3%)	30.5	41.9% (3%)	3.8
3	30.9% (2.7%)	99.1	29.2% (2.6%)	0	29.5% (2.7%)	0.9
4	26.1% (2.3%)	66.6	26% (2.3%)	27.2	26% (2.3%)	6.2
5	57.7% (4.1%)	64.6	57.6% (4.1%)	32.1	57.6% (4.1%)	3.3
6	46.2% (3.8%)	99.4	43.6% (3.8%)	0	44.2% (3.8%)	0.6

The percentage of variance explained is denoted by VE. The quantitative trait is assumed to follow a normal distribution with nine different means conditionined on the combination of joint genotypes at the two QTLs. The allele frequencies at the two QTLs are 0.1 and 0.1. The conditional means of Y given genotypes at the two QTLs are expressed in the following form: $E(Y|G_iH_j) = \mu_{ij} = a_i + b_j$, where $a_1 = \alpha_1, a_2 = \beta_1, a_3 = -\alpha_1$, and $b_1 = \alpha_2, b_2 = \beta_2, b_3 = -\alpha_2$. Here G_i and H_j $(i,j \in (1,2,3))$ denote the pair of genotypes at the two QTLs. The parameters $\{(\alpha_1,\beta_1),(\alpha_2,\beta_2)\}$ are chosen as $\{(1.0,0.0),(1.0,0.0)\}$, $\{(2.0,0.0),(2.0,0.0)\}$, $\{(2.0,0.0),(1.0,0.0)\}$, $\{(1.0,0.5),(1.0,0.5)\}$, $\{(2.0,1.0),(2.0,1.0)\}$, $\{(2.0,1.0),(1.0,0.5)\}$, for quantitative trait models 1, 2, 3, 4, 5, 6, respectively.

higher risk at the first disease locus compared to the second one.

We present the comparative performance of different measures of allele dosage scores in tables 1, 2 and 3 for a binary trait corresponding to disease models 1–18 outlined in table 4. We observe that the allele dosage score based on log OR performs overall better than the other two measures. However in many scenarios, the mean AUC (rounded to second decimal point) is the same for all measures. When ADS_{OR} performs better for a disease model, the improvement in AUC is marginal. For example, for disease models 2, 4 and 6, ADS_{OR} produced 1% higher AUC compared to the other two measures (table 1). But for all of the other disease models considered in tables 1, 2 and 3, the three ADS measures produced very similar mean AUCs.

Table 7. Mean variance explained and the percentage of times maximum variance explained of the trait by each measure of allele dosage score for quantitative trait models 7, 8, 9, 10, 11 and 12.

	ADS_{Beta}		ADS_{PV}		ADS_{VE}	
Quantitative trait models	Mean VE	Per cent of times maximum VE	Mean VE	Per cent of times maximum VE	Mean VE	Per cent of times maximum VE
7	32.6% (2.1%)	100	31.2% (2.1%)	0	31.4% (2.1%)	0
8	64.3% (3.4%)	100	62.7% (3.4%)	0	62.7% (3.4%)	0
9	50.8% (3.1%)	40.4	50.5% (3.4%)	11.7	50.9% (3.3%)	47.9
10	45.1% (2.5%)	100	43.6% (2.5%)	0	44.1% (2.5%)	0
11	73.1% (3.8%)	100	72.1% (3.8%)	0	72.1% (3.8%)	0
12	64.6% (4.2%)	69.3	64.2% (4.4%)	30.7	64.2% (4.4%)	0

The percentage of variance explained is denoted by VE. The quantitative trait is assumed to follow a normal distribution with nine different means conditioned on the combination of joint genotypes at the two QTLs. The allele frequencies at the two QTLs are 0.1 and 0.2. The conditional means of Y given genotypes at the two QTLs are expressed in the following form: $E(Y|G_iH_j) = \mu_{ij} = a_i + b_j$, where $a_1 = \alpha_1, a_2 = \beta_1, a_3 = -\alpha_1$, and $b_1 = \alpha_2, b_2 = \beta_2, b_3 = -\alpha_2$. Here G_i and H_j $(i,j \in (1,2,3))$ denote the pair of genotypes at the two QTLs. The parameters $\{(\alpha_1,\beta_1),(\alpha_2,\beta_2)\}$ are chosen as $\{(1.0,0.0),(1.0,0.0)\}$, $\{(2.0,0.0),(2.0,0.0)\}$, $\{(2.0,0.0),(1.0,0.0)\}$, $\{(1.0,0.5),(1.0,0.5)\}$, $\{(2.0,1.0),(2.0,1.0)\}$, $\{(2.0,1.0),(1.0,0.5)\}$, for quantitative trait models 7, 8, 9, 10, 11, 12, respectively.

Table 8. Mean variance explained and the percentage of times maximum variance explained of the trait by each measure of allele dosage score for quantitative trait models 13, 14, 15, 16, 17 and 18.

	ADS_{Beta}		ADS_{PV}		ADS_{VE}	
Quantitative trait models	Mean VE	Per cent of times maximum VE	Mean VE	Per cent of times maximum VE	Mean VE	Per cent of times maximum VE
13	15.4% (1.7%)	63.6	15.3% (1.7%)	28.9	15.3% (1.7%)	7.5
14	42% (3%)	65.7	41.9% (3%)	30.5	41.9% (3%)	3.8
15	30.9% (2.7%)	99.1	29.2% (2.6%)	0	29.5% (2.7%)	0.9
16	26.1% (2.3%)	66.6	26% (2.3%)	27.2	26% (2.3%)	6.2
17	57.7% (4.1%)	64.6	57.6% (4.1%)	32.1	57.6% (4.1%)	3.3
18	46.2% (3.8%)	99.4	43.6% (3.8%)	0	44.2% (3.8%)	0.6

The percentage of variance explained is denoted by VE. The quantitative trait is assumed to follow a chi-square distribution with nine different means conditioned on the combination of joint genotypes at the two QTLs. The allele frequencies at the two QTLs are 0.1 and 0.1. The conditional means of Y given genotypes at the two QTLs are expressed in the following form: $E(Y|G_iH_j) = \mu_{ij} = a_i + b_j$, where $a_1 = \alpha_1, a_2 = \beta_1, a_3 = -\alpha_1$, and $b_1 = \alpha_2, b_2 = \beta_2, b_3 = -\alpha_2$. Here G_i and H_j $(i,j \in (1,2,3))$ denote the pair of genotypes at the two QTLs. The parameters $\{(\alpha_1,\beta_1),(\alpha_2,\beta_2)\}$ are chosen as $\{(1.0,0.0),(1.0,0.0)\}$, $\{(2.0,0.0),(2.0,0.0)\}$, $\{(2.0,0.0),(1.0,0.0)\}$, $\{(1.0,0.5),(1.0,0.5)\}$, $\{(2.0,1.0),(2.0,1.0)\}$, $\{(2.0,1.0),(1.0,0.5)\}$, for quantitative trait models 13, 14, 15, 16, 17, 18, respectively.

In table 5, we present the results on comparative performance of the three measures under a logistic disease model. We considered 20 different scenarios in which the OR for five risk SNPs were simulated at random from Uniform(1.1, 1.5). We again observe that in some cases, ADS_{OR} produced marginally higher (1%) mean AUCs than the other two measures. Thus, for a binary trait, ADS_{OR} performs marginally better than the other two approaches.

For a quantitative trait, we consider different choices of (α_1,β_1) and (α_2,β_2) to vary the effect sizes of the QTLs $(a_1 = \alpha_1, a_2 = \beta_1, a_3 = -\alpha_1,$ and $b_1 = \alpha_2, b_2 = \beta_2, b_3 = -\alpha_2)$. The results obtained from the simulations are presented in tables 6, 7, 8, 9 and 10. We observe that, for a quantitative trait, the measure based on the standardized estimate of the slope coefficient (ADS_{Beta}) performs overall better than the other two measures. For example, for quantitative trait models 3, 6 and 12 (tables 6 & 7), ADS_{Beta} produced 1.7%, 2.6% and 2.4% higher proportion of variance explained compared to ADS_{PV}, respectively. However, the improvement is very marginal in some cases, e.g., for quantitative trait models 1, 2, 4 and 5 (table 6). When we considered an extra epistasis term in the phenotype model, ADS_{Beta} performed marginally better (table 10). Hence, overall, ADS_{Beta} performs marginally better than the other two approaches.

Table 9. Mean variance explained and the percentage of times maximum variance explained of the trait by each measure of allele dosage score for quantitative trait models 19, 20, 21, 22, 23 and 24.

	ADS_{Beta}		ADS_{PV}		ADS_{VE}	
Quantitative trait models	Mean VE	Per cent of times maximum VE	Mean VE	Per cent of times maximum VE	Mean VE	Per cent of times maximum VE
19	19.7% (1.8%)	100	18.9% (1.8%)	0	19% (1.8%)	0
20	48.6% (3%)	100	46.4% (2.9%)	0	47.1% (3%)	0
21	33.7% (2.6%)	36.7	33.8% (2.8%)	15.9	33.9% (2.8%)	47.4
22	29.9% (2.3%)	99.9	29% (2.2%)	0.1	29.1% (2.2%)	0
23	60.5% (3.6%)	100	58.6% (3.6%)	0	59.3% (3.6%)	0
24	48.4% (3.5%)	56	47.8% (3.8%)	3.5	48.3% (3.7%)	40.5

The percentage of variance explained is denoted by VE. The quantitative trait is assumed to follow a chi-square distribution with nine different means conditioned on the combination of joint genotypes at the two QTLs. The allele frequencies at the two QTLs are 0.1 and 0.2. The conditional means of Y given genotypes at the two QTLs are expressed in the following form: $E(Y|G_iH_j) = \mu_{ij} = a_i + b_j$, where $a_1 = \alpha_1, a_2 = \beta_1, a_3 = -\alpha_1$, and $b_1 = \alpha_2, b_2 = \beta_2, b_3 = -\alpha_2$. Here G_i and H_j $(i,j \in (1,2,3))$ denote the pair of genotypes at the two QTLs. The parameters $\{(\alpha_1,\beta_1),(\alpha_2,\beta_2)\}$ are chosen as $\{(1.0,0.0),(1.0,0.0)\}$, $\{(2.0,0.0),(2.0,0.0)\}$, $\{(2.0,0.0),(1.0,0.0)\}$, $\{(1.0,0.5),(1.0,0.5)\}$, $\{(2.0,1.0),(2.0,1.0)\}$, $\{(2.0,1.0),(1.0,0.5)\}$, for quantitative trait models 19, 20, 21, 22, 23, 24, respectively.

Table 10. Mean variance explained and the percentage of times maximum variance explained of the trait by each measure of allele dosage score for quantitative trait models 25, 26, 27, 28, 29 and 30.

	ADS_{Beta}		ADS_{PV}		ADS_{VE}	
Quantitative trait models	Mean VE	Per cent of times maximum VE	Mean VE	Per cent of times maximum VE	Mean VE	Per cent of times maximum VE
25	11.4% (1.3%)	66.1	11.4% (1.3%)	22.1	11.4% (1.3%)	11.8
26	4.9% (1.1%)	62.9	4.9% (1.1%)	12.6	4.9% (1.1%)	24.5
27	20.6% (1.9%)	96.6	20.2% (1.9%)	2.8	20.2% (1.9%)	0.6
28	22.1% (1.8%)	57.8	22.1% (1.8%)	35	22.1% (1.8%)	7.2
29	13.3% (1.8%)	56.9	13.3% (1.8%)	31.4	13.3% (1.8%)	11.7
30	35% (2.8%)	99	34% (2.8%)	0	34.1% (2.9%)	1

The percentage of variance explained is denoted by VE. The quantitative trait is assumed to follow a normal distribution with nine different means conditioned on the combination of joint genotypes at the two QTLs. The allele frequencies at the two QTLs are 0.1 and 0.1. The conditional means of Y given genotypes at the two QTLs are expressed in the following form: $E(Y|G_iH_j) = \mu_{ij} = a_i + b_j + \frac{a_i b_j}{2}$, where $a_1 = \alpha_1, a_2 = \beta_1, a_3 = -\alpha_1$, and $b_1 = \alpha_2, b_2 = \beta_2, b_3 = -\alpha_2$. We considered an additional epistasis term. Here, G_i and H_j $(i,j \in (1,2,3))$ denote the pair of genotypes at the two QTLs. The parameters $\{(\alpha_1,\beta_1),(\alpha_2,\beta_2)\}$ are chosen as $\{(1.0,0.0),(1.0,0.0)\}$, $\{(2.0,0.0),(2.0,0.0)\}$, $\{(2.0,0.0),(1.0,0.0)\}$, $\{(1.0,0.5),(1.0,0.5)\}$, $\{(2.0,1.0),(2.0,1.0)\}$, $\{(2.0,1.0),(1.0,0.5)\}$, for quantitative trait models 25, 26, 27, 28, 29, 30, respectively.

Application using real data

In a study on T2DM, Ramya *et al.* (2013) analysed eight variants in *Adiponectin* gene using data on T2DM individuals and unrelated nonglucose tolerant (NGT) individuals selected from the Chennai Urban and Rural Epidemiology Study (CURES) in south India. Significant evidence of association with T2DM was obtained at five SNPs: rs17300539, rs266729, rs822393, rs1501299, and rs3774261. After initial filtering of the original data to remove individuals with missing genotype information at any of these five SNPs, we obtained 299 case and 582 control individuals. We assume the prevalence of T2DM in the selected Indian population to be 0.1. Thus, we consider a training sample comprising 50 cases and 450 controls in our computations to maintain a reasonably appropriate representation of cases and controls. We evaluate the different measures of the allele dosage score with respect to AUC while estimating the risk of T2DM as a binary trait. For FBS levels, we compare the measures of allele dosage score with respect to the proportion of variance of FBS explained. For FBS level, we consider one half of the individuals as the training sample. For both of T2DM and FBS levels, we consider the remaining individuals as the test sample.

Several studies have shown that body mass index (BMI) is a potential confounder in the association analyses of T2DM since there are common SNPs that regulate both the phenotypes. Hence, while estimating the OR, *P*-values corresponding to the tests for association and the proportion of variance of the trait explained by each of the selected SNPs required for computing the weights involved in the different measures of allele dosage scores for T2DM, we adjust for the BMI levels along with age. We also do an identical adjustment when we predict T2DM risk of the individuals in the test sample. Similarly in the analysis of FBS level, we adjust for BMI level and age.

For T2DM, we randomly select a set of 50 individuals from the set of 299 case individuals and 450 individuals from the pool of 582 control individuals. The combined group of these randomly selected case and control individuals was treated as the training sample and the set of remaining individuals was used as the test sample. We estimate the weights involved in allele dosage scores based on the training sample. We predict the risk of individuals in test sample based on the allele dosage score while adjusting for BMI and age. We estimate the AUC corresponding to each measure of allele dosage score based on the test sample.

For the FBS level as the quantitative trait, we estimate the weights involved in the allele dosage scores based on the training sample (half of the sample randomly selected) while adjusting for BMI and age. We estimate the proportion of variance of FBS level explained by each measure of allele dosage score based on the test sample while adjusting for BMI and age.

We estimated the AUCs for the three measures of ADS while predicting the risk of T2DM, and the proportion of variance explained for FBS level. We repeated the analysis for 1000 random choices of the training and test samples. For T2DM, the mean AUC obtained by ADS_{Beta}, ADS_{PV}, ADS_{VE} were 0.65, 0.65 and 0.64, respectively. Thus, ADS_{Beta}, ADS_{PV} performed similarly and produced 1% better AUC than ADS_{VE}. For FBS level, the percentage of variance explained by ADS_{Beta}, ADS_{PV}, ADS_{VE} were 5.9%, 5.8% and 5.7%, respectively. Hence, all the measures performed comparably for the quantitative trait.

Discussion

In this article, we have compared some intuitive choices of measures of allele dosage scores to integrate associated SNPs identified in genomewide studies to predict the risk of a clinical outcome, or estimate the genetic variance of a quantitative trait explained by the associated SNPs. We have explored the utility and relevance of these measures in case of both binary and quantitative traits using simulations and real data analysis. We have theoretically shown that the genetic risk of a disease should not be predicted based on usual association study designs comprising equal number of cases and controls.

Our simulation study suggests that for a binary trait, the allele dosage score defined based on log OR performs marginally better than the other two measures with respect to marginally better AUC. Therefore, ADS_{OR} should be used in general, irrespective of the underlying true disease models. For a quantitative trait, ADS_{Beta} performs consistently better than the other two measures with a marginal improvement in the proportion of the variance of the phenotype explained due to the allele dosage scores. Thus, ADS_{Beta} should be the preferred choice for a quantitative trait irrespective of the phenotype model. The computations based on real data analyses also validate these findings to a large extent. We postulate that the relatively better predictions obtained using ADS_{OR} in the context of binary traits and ADS_{Beta} in the context of quantitative traits can be attributed primarily to the fact that the estimated OR and β are the maximum likelihood estimators of the incremental disease risk and increase in quantitative trait values, respectively, due to each additional risk allele.

Appendix

Risk prediction for a binary end-point trait

The popular case–control study design usually involves a prefixed number of cases and controls. We demonstrate that such a design is inappropriate for estimating risks conferred by individual SNPs. The prevalence of a disease is a function of the allele frequencies and the penetrances at the different causal loci. Since the prevalence of a disease in a population is usually much lower than 50% (or the proportion of cases in GWAS data), there is a gross over-representation of cases in the usual study design comprising equal number of cases and controls (or with substantially larger case control ratio compared to the population) leading to overestimation of the penetrances. On the other hand, a large sample of individuals chosen at random from a population irrespective of their affection status is expected to carry proper information on the disease prevalence in the population.

For a disease trait, define a variable Y for an individual, such that $Y = 1$, if the individual is a case, and $Y = 0$ if the individual is a control. We assume that the disease locus is biallelic with alleles D and d having frequencies p and $(1-p)$, respectively in the population. Suppose we have genotype data at the disease locus on $2n$ individuals comprising n_1 cases and n_2 $(= 2n - n_1)$ controls. The penetrances at the disease locus are defined as: $f_2 = P(\text{case}|DD)$, $f_1 = P(\text{case}|Dd)$, $f_0 = P(\text{case}|dd)$. The overall disease prevalence is given by: $P(\text{case}) = f_2p^2 + 2f_1p(1-p) + f_0(1-p)^2$. Using a logistic link function between the disease status and SNP genotype, we can estimate the penetrance parameters from the data.

$P(DD|\text{case}) = \frac{f_2p^2}{P(\text{case})}$, $P(Dd|\text{case}) = \frac{2f_1p(1-p)}{P(\text{case})}$, $P(dd|\text{case}) = \frac{f_0(1-p)^2}{P(\text{case})}$, and, $P(DD|\text{control}) = \frac{(1-f_2)p^2}{P(\text{control})}$, $P(Dd|\text{control}) = \frac{2(1-f_1)p(1-p)}{P(\text{control})}$, $P(dd|\text{control}) = \frac{(1-f_0)(1-p)^2}{P(\text{control})}$, where

$P(\text{control}) = 1 - P(\text{case})$. Let n_{case}^{DD} and $n_{control}^{DD}$ denote the number of case and control individuals who have the genotype DD, respectively. Similarly, n_{case}^{Dd}, n_{case}^{dd}, $n_{control}^{Dd}$, $n_{control}^{dd}$ are defined. Hence, $E(n_{case}^{G}) = n_1 \times P(G|\text{case})$ and $E(n_{control}^{G}) = n_2 \times P(G|\text{control})$, $G = DD, Dd, dd$. Since for a given G, n_{case}^{G}/n_1 is a consistent estimator of $P(G|\text{case})$, we assume that, $n_{case}^{G} \approx n_1 \times P(G|\text{case})$, for large n_1. Using the same argument, for a given G, we assume that $n_{control}^{G} \approx n_2 \times P(G|\text{control})$, for large n_2.

Let X denote the number of D alleles in a genotype and hence, $X = 0, 1, 2$. We model the probability of a case conditioned on the genotype at the disease locus via the logistic link as follows: $P(\text{case}|X = x) = \frac{\exp\{\alpha+\beta x\}}{1+\exp\{\alpha+\beta x\}}$. So, $P(\text{control}|X = x) = \frac{1}{1+\exp\{\alpha+\beta x\}}$.

Following the above discussion, for a given choice of n_1, n_2, and f_2, f_1, f_0, and p, we assume the values of n_{case}^{G} and $n_{control}^{G}$ to be $n_1 P(G|\text{case})$ and $n_2 P(G|\text{control})$, respectively, $G = DD, Dd, dd$. Suppose, we model the likelihood L of the genotype data of the cases and controls: $n_{case}^{G}(= n_1^{G})$ and $n_{control}^{G}(= n_2^{G})$, $G = DD, Dd, dd$, based on the logistic model presented above. Then the log-likelihood function denoted by $\log(L)$ is given by:

$$\begin{aligned}\log(L) =& n_1^{DD}(\alpha + 2\beta) + n_1^{Dd}(\alpha + \beta) + n_1^{dd}\alpha \\ &- (n_1^{DD} + n_2^{DD})log(1 + exp\{\alpha + 2\beta\}) \\ &- (n_1^{Dd} + n_2^{Dd})log(1 + exp\{\alpha + \beta\}) \\ &- (n_1^{dd} + n_2^{dd})log(1 + exp\{\alpha\}).\end{aligned}$$

For ease of exposition, we define the following quantities: $\text{prob}_2 = \frac{\exp\{\alpha+2\beta\}}{1+\exp\{\alpha+2\beta\}}$, $\text{prob}_1 = \frac{\exp\{\alpha+\beta\}}{1+\exp\{\alpha+\beta\}}$, $\text{prob}_0 = \frac{\exp\{\alpha\}}{1+\exp\{\alpha\}}$. Then the 1st and 2nd order partial derivatives of the log data likelihood are given by:

$$\begin{aligned}\frac{\partial \log(L)}{\partial\alpha} &= (n_1^{DD} + n_1^{Dd} + n_1^{dd}) - (n_1^{DD} + n_2^{DD})\ \text{prob}_2 \\ &\quad - (n_1^{Dd} + n_2^{Dd})\ \text{prob}_1 - (n_1^{dd} + n_2^{dd})\ \text{prob}_0 \\ \frac{\partial \log(L)}{\partial\beta} &= 2n_1^{DD} + n_1^{Dd} - 2(n_1^{DD} + n_2^{DD})\ \text{prob}_2 \\ &\quad - (n_1^{Dd} + n_2^{Dd})\ \text{prob}_1 \\ \frac{\partial^2 \log(L)}{\partial\alpha^2} &= -(n_1^{DD} + n_2^{DD})\ \text{prob}_2(1 - \text{prob}_2) \\ &\quad - (n_1^{Dd} + n_2^{Dd})\ \text{prob}_1(1 - \text{prob}_1) \\ &\quad - (n_1^{dd} + n_2^{dd})\ \text{prob}_0(1 - \text{prob}_0) \\ \frac{\partial^2 \log(L)}{\partial\beta^2} &= -2(n_1^{DD} + n_2^{DD})\ \text{prob}_2(1 - \text{prob}_2) \\ &\quad - (n_1^{Dd} + n_2^{Dd})\ \text{prob}_1(1 - \text{prob}_1).\end{aligned}$$

Using the above equations in the Fisher's scoring method, we obtain m.l.e. of α and β. Alternatively, we could also use the Newton Raphson method or the iteratively reweighted least squares method to obtain the m.l.e. Based on the m.l.e. of α and β, we estimate the penetrance parameters f_2, f_1, f_0, as $P(\text{case}|X = 2), P(\text{case}|X = 1), P(\text{case}|X = 0)$, respectively, using the logistic model discussed above.

For a given choice of n_1, n_2, and f_2, f_1, f_0, and p, (and hence, n_1^{G}, n_2^{G}, $G = DD, Dd, dd$), it is expected that the estimated penetrances would be close to the true choice of f_2, f_1, f_0. For example, we consider a choice of the penetrance parameters: $f_2 = 0.1, f_1 = 0.05, f_0 = 0.01$, and $p = 0.1$, that induces an overall prevalence of 0.018. Suppose we consider a total sample of 10,000 individuals. If we consider a design with equal number of cases and controls (i.e., $n_1 = n_2 = 5000$) and estimate f_2, f_1, f_0 using the above model, the estimates are found to be $f_2 = 0.93, f_1 = 0.73, f_0 = 0.36$. Hence, it is clear that the true penetrances are grossly overestimated as pointed out earlier. On the other hand, if we consider the number of cases to be proportional to the overall prevalence (i.e., $n_1 = 181$, $n_2 = 9819$) we obtain the estimates of the penetrances to be $f_2 = 0.16, f_1 = 0.04, f_0 = 0.01$. Thus, these estimates are much closer to the true penetrances compared to the previous sampling design comprising equal number of cases and controls.

Acknowledgements

The authors are grateful to Prof. V. Mohan and Dr Radha Venkatesan of Madras Diabetes Research Foundation (MDRF) for providing access to an apriori analysed portion of the CURES data through a mutual collaboration with SG.

References

Chauhan G., Spurgeon C. J., Tabassum R., Bhaskar S., Kulkarni S. R., Mahajan A. *et al*. 2010 Impact of common variants of PPARG, KCNJ11, TCF7L2, SLC30A8, HHEX, CDKN2A, IGF2BP2, and CDKAL1 on the risk of type 2 diabetes in 5,164 Indians. *Diabetes* **59**, 2068–2074.

Consortium I. S. 2009 Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder. *Nature* **460**, 748.

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

Duncan L., Shen H., Gelaye B., Meijsen J., Ressler K., and Feldman M. 2019 Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 1–9.

Hachiya T., Kamatani Y., Takahashi A., Hata J., Furukawa R., Shiwa Y. *et al*. 2017 Genetic predisposition to ischemic stroke: a polygenic risk score. *Stroke.* **48**, 253–258

Kendler K. S. 2016 The schizophrenia polygenic risk score: to what does it predispose in adolescence?. *JAMA Psychiat*. **73**, 193–194.

Khera A. V., Chaffin M., Aragam K. G., Haas M. E., Roselli C., Choi S. H. *et al*. 2018 Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224.

Martin A. R., Kanai M., Kamatani Y., Okada Y., Neale B. M. and Daly M. J. 2019 Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591.

Mega J. L., Stitziel N. O., Smith J. G., Chasman D. I., Caulfield M. J., and Devlin J. J. 2015 Genetic risk, coronary heart disease

events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet* **385**, 2264–2271.

Peyrot W. J., Milaneschi Y., Abdellaoui A., Sullivan P. F., Hottenga J. J., Boomsma D. I. and Penninx B. W. 2014 Effect of polygenic risk scores on depression in childhood trauma. *Br. J. Psychiat.* **205**, 113–119.

Prentice R. L. and Pyke R. 1979 Logistic disease incidence models and case-control studies. *Biometrika*. **66**, 403–411.

Power R. A., Steinberg S., Bjornsdottir G., Rietveld C. A., Abdellaoui A., Nivard M. M. *et al*. 2015 Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat. Neurosci.* **18**, 953–955.

Ramya K., Ayyappa K. A., Ghosh S., Mohan V. and Radha V. 2013 Genetic association of ADIPOQ gene variants with type 2 diabetes, obesity and serum adiponectin levels in south Indian population. *Gene*. **532**, 253–262.

Richardson T. G., Harrison S., Hemani G. and Smith G. D. 2019 An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife.* **8**, e43657.

Torkamani A., Wineinger N. E. and Topol E. J. 2018 The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590.

Wray N. R., Goddard M. E. and Visscher P. M. 2007 Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528.

Corresponding editor: Shrish Tiwari