European Journal of Human Genetics (2010) 18, 485–489

www.nature.com/ejhg

ARTICLE

Evaluation of the discriminative accuracy of genomic profiling in the prediction of common complex diseases

Ramal Moonesinghe*[,1], Tiebin Liu[2] and Muin J Khoury[2]

Genetic testing for susceptibility to common diseases based on a combination of genetic markers may be needed because the effect size associated with each genetic marker is small. Whether or not a genome profile based on a combination of markers could yield a useful test can be evaluated by assessing the discriminative accuracy. The authors present a simple method to calculate the clinical discriminative accuracy of a genomic profile when the relative risk and genotype frequency of each genotype are known. In addition, the clinical discriminative accuracy of a genetic test is presented for given values of the heritability and prevalence of the disease and for the population-attributable fraction of the combined genetic markers. For given values of relative risk and genotype frequency, the discriminative accuracy increases with increasing heritability but declines with increasing prevalence of the disease. For a given value of population-attributable fraction, the discriminative accuracy increases with increasing relative risks, but declines with increasing genotype frequency. On the basis of population-attributable fraction and estimates of heritability of disease, the number of risk genotypes required to have a reasonable clinical discriminative accuracy is much higher than the genome profiles available at present.
European Journal of Human Genetics (2010) **18,** 485–489; doi:10.1038/ejhg.2009.209; published online 25 November 2009

Keywords: genomic profiles; discriminative accuracy; heritability

INTRODUCTION

Testing for susceptibility to common diseases may provide a unique opportunity for disease prevention. Nevertheless, because of the complex nature of common diseases with multiple genetic and environmental risk factors, genetic testing for these diseases may require bundling multiple genetic markers because the risk associated with each genetic marker is very small. Many authors have discussed the discriminative accuracy of combining several risk factors as a screening test.[1–3] For complex diseases, most of the genetic associations have odds ratios ranging from 1.1 to 1.5, and any single polymorphism accounts for only 1–8% of the overall disease risk in the population.[4] The concept of using genetic variants at multiple loci simultaneously is known as genomic profiling.[5]

A first evaluation on whether a combination of genetic risk factors can potentially yield a useful predictive test is indicated by the discriminative accuracy. The discriminative accuracy is generally indicated by the area under the receiver-operating characteristic (ROC) curve, which originally developed to evaluate the performance of a single test is a method of describing the intrinsic accuracy of a test apart from the decision thresholds. An ROC curve is a plot of a test's sensitivity *vs* its false-positive rate or (1-specificity) for every possible cutoff value of the continuous test result. ROC curves and their characteristics have been described in many papers.[6–8]

Janssens *et al.*[9] investigated the impact of the frequencies of individual risk genotypes on the clinical validity of genomic profiling. They evaluated the clinical discriminative accuracy (area under the ROC curve, AUC) and disease risks for the simultaneous testing of 40 independent susceptibility genetic variants. Their results are based on varying the genotype frequency and odds ratio and evaluating the corresponding AUC using separate simulation scenarios. In this paper, we provide a theoretical framework for genomic profiling and derive equations for ROC curves and AUC, so that any given scenario can be studied without conducting simulation studies. We start with formulas for identical genetic markers and extend our method to evaluate clinical discriminative accuracy for testing for multiple genetic variants with different relative risks and genotype frequency. We compare our results with the results obtained by simulation and show that almost identical. We calculate the discriminative accuracy of predictive testing for multiple genetic variants using an example of five SNPs associated with prostate cancer. We chose this example because a test had been made available based on these five SNPs. In addition, we also present formulas to calculate the AUC based on the heritability of diseases and the population-attributable fractions (PAFs) of the genetic variants in the genomic profile.

METHODS

Suppose that the population at risk is exposed to a level X_i of the ith genetic variant (X_i can assume only 1 or 0 depending on the presence or absence of the ith genetic variant).

Let $G_1, G_2, \ldots, G_k$ and $R_1, R_2, \ldots, R_k$ be the prevalence and the relative risks for the k-genetic variants and are assumed to be known. Assuming that the exposure variables, X_i, corresponding to the k-genetic variants are independent, the joint distribution of the k exposure variables is given by

$$f(X_1, X_2, \ldots, X_k) = G_1^{X_1}(1-G_1)^{(1-X_1)} \ldots G_k^{X_k}(1-G_k^{X_k}) = \prod_{i=1}^{k} G_i^{X_i}(1-G_i)^{(1-X_i)}$$

[1]Office of Minority Health and Health Disparities, Centers for Disease Control and Prevention, Atlanta, GA, USA; [2]Office of Public Health Genomics, Coordinating Center for Health Promotion, Centers for Disease Control and Prevention, Atlanta, GA, USA
*Correspondence: Dr R Moonesinghe, Office of Minority Health and health Disparities, Centers for Disease Control and Prevention, Mailstop E-67, 1600 Clifton Road, NE, Atlanta, GA 30333, USA. Tel: +1 404 498 2342; Fax: +1 770 488 8336; E-mail: rmoonesinghe@cdc.gov
Received 20 May 2009; revised 19 October 2009; accepted 22 October 2009; published online 25 November 2009

If U_1, U_2,…,U_k, denote the exposure variables (U_i can assume only 0 or 1) among cases for the k-genetic variants, the joint distribution of U_1, U_2,…,U_k, under the assumption of a multiplicative risk model is given by[10]

$$g(U_1, U_2, \ldots, U_k) = \prod_{i=1}^{k} (G_i^*)^{U_i}(1 - G_i^*)^{(1-U_i)},$$

where $G_i^* = \frac{R_i G_i}{R_i G_i + (1-G_i)}$.

Calculation of AUC for *k* identical risk genotypes

To study the effect of genotype frequencies and relative risks, we first assumed that all genetic markers had the same effect size (R_i=R, I=1, 2, … , k) and the same genotype frequency (G_i=G, i=1, 2, … , k). We also assumed that X_i's are exposure variables for controls.[10]

Let $X=\sum_{i=1}^{k} X_i$ and $U=\sum_{i=1}^{k} U_i$. X and U represent the number of risk genotypes for controls and cases, respectively. The distribution of X for controls has a binomial distribution with parameter G and the distribution of U for cases has a binomial distribution with parameter G^*. For large k values, the distributions of X and U can be approximated by normal distributions. Suppose the cumulative distributions of X and U are given by F and H, respectively. F is a normal distribution with mean kG and variance kG (1–G) and H is a normal distribution with mean kG^* and variance $kG^*(1–G^*)$. The ROC curve can then be expressed as $\mathrm{ROC}(t) = \bar{\mathrm{H}}(\bar{\mathrm{F}}^{-1}(t))$ for $0 < t < 1$, where $\bar{\mathrm{F}}$ and $\bar{\mathrm{H}}$ are the survival functions of F and H, respectively.[7] The survival function $\bar{\mathrm{F}}$ for controls is given by $\bar{\mathrm{F}}(t) = 1 - \Phi\left[\frac{t-kG}{\sqrt{kG(1-G)}}\right]$ and the survival function $\bar{\mathrm{H}}$ for cases is given by $\bar{\mathrm{H}}(t) = 1 - \Phi\left[\frac{t-kG^*}{\sqrt{kG^*(1-G^*)}}\right]$, where Φ is the cumulative distribution function of the normal distribution. The ROC curve can then be expressed as $\mathrm{ROC}(t) = \bar{\mathrm{H}}(\bar{\mathrm{F}}^{-1}(t)) = 1 - \Phi\left[\frac{\sqrt{kG(1-G)}\Phi^{-1}(1-t)+k(G-G^*)}{\sqrt{kG^*(1-G^*)}}\right]$, for $0 < t < 1$, and the AUC is given by $\Phi\left[\frac{k(G^*-G)}{\sqrt{kG^*(1-G^*)+kG(1-G)}}\right]$.

Heritability and PAF

The analytical expressions given above are valid for any number of markers (k). However, it is possible to investigate the relationship between the relative risk of genetic markers and the number of markers that contribute to risk of disease based on known disease prevalence and heritability or known PAF. Let p be the disease prevalence and h^2 the heritability of the disease, the number of markers that contribute to risk of disease is then given by[11] (Appendix A):

$$k = [\log\{h^2(1-p)+p\} - \log p]/[\log\{R^2G+(1-G)\} - 2\log\{RG+(1-G)\}].$$

Similarly, the number of markers that contribute to risk of disease when PAF is known is given by[12]:

$$k = \frac{-\log(1-\mathrm{PAF})}{\log[GR+(1-G)]}$$

Calculation of AUC for *k* risk genotypes with different relative risks and genotype frequency

Next, we consider the scenario when genotype frequency and effect sizes are different for different genetic variants. Let **X** be a vector of values of k different markers obtained from a randomly picked patient in the controls and let **U** be a similar vector of a randomly picked patient in the cases. Consider the linear combinations, $V = \sum_{i=1}^{k} a_i X_i$ and $W = \sum_{i=1}^{k} a_i U_i$. Su and Liu[13] showed that Fisher's linear discriminant function (LDF) provides a linear combination of markers to maximize the sensitivity over the entire specificity range uniformly under the multivariate normal distribution model with proportional covariance matrices. They also provided a solution of the best linear combination of markers in the sense that the AUC of this combination is maximized among all possible linear combinations. The LDF is frequently applied to binary variables.[14–16] Almost all the results suggest that LDF can be recommended for binary variables because of its expected stability as the number of variables increases. Furthermore, as shown in many studies that have been conducted, the losses incurred by LDF under nonoptimal conditions compared with other procedures are small enough not to be of any practical importance.[17] Applying the results of Su and Liu[13] to our binary variables, the optimal linear combination is given by the coefficients:

$$a_i = \frac{G_i^* - G_i}{G_i^*(1-G_i^*)+G_i(1-G_i)},$$

i=1, 2, … , k, and the AUC of the optimal combination is given by

$$\mathrm{A} = \Phi\left(\sqrt{\sum_{i=1}^{k} \frac{(G_i^* - G_i)^2}{G_i^*(1-G_i^*)+G_i(1-G_i)}}\right).$$

Note that when the genotype frequency and relative risks are identical, this expression is identical to the AUC derived for this situation previously.

RESULTS

Figure 1 gives the AUC when relative risks are 1.1, 1.3 and 1.5, and genotype frequencies are 0.1, 0.2 and 0.3, for 40 and 80 identical markers. As expected, the lowest AUC, 0.55, corresponds to the lowest relative risk, 1.1, and lowest genotype frequency, 0.1, for 40 markers. When relative risk is 1.1 for 40 markers, the increase in AUC obtained by increasing the genotype frequency from 0.1 to 0.3 is only 0.03 or 5%; when relative risk is 1.5, the increase in AUC is 0.09 or 12.5%. Similarly, when genotype frequency is 0.1, for 40 markers, the increase in AUC obtained by increasing the relative risk from 1.1 to 1.5 is 0.17 or 31%; when genotype frequency is 0.3, the increase in AUC is 0.23 or 40%. Overall, the AUC increases when the number of markers are doubled to 80 with the smallest increase corresponding to the lowest values of relative risk and genotype frequency and the largest increase corresponding to the highest values of relative risk and genotype frequency. Genotypes with low relative risk (around 1.1) and genotype frequency (around 0.1) require about 1000 markers in the genomic profile to have a reasonable discriminative power (AUC around 0.74).

When the relative risk is 1.1, genotype frequency, 0.1, disease prevalence, 5%, and heritability, 5%, the number of markers that contribute to risk of disease is 758 (AUC=0.71); if we increase heritability to 10%, the number of markers that contribute to risk of disease is 1208 (AUC=0.76). With heritability at 5%, when the disease prevalence is increased to 10%, the number of markers decline to 422 (AUC=0.66). When heritability and relative risk are also increased to 10% and 1.5, respectively, the number of markers that contribute to disease decline to 32 (AUC=0.70); when genotype frequency is increased to 0.3, the number of markers decline further to 17 (AUC=0.71). Interestingly, AUC as a function of heritability and disease prevalence remains approximately same for the ranges of relative risks (1.1–1.5) and genotype frequencies (0.1–0.3) considered. For example, when heritability is 5% and disease prevalence is 5%, the AUC remains around 0.71; when disease prevalence is increased to 10%, AUC declines to 0.66. When heritability is 10% and disease prevalence is 5%, the AUC is 0.76; increasing disease prevalence to 10% results in a decline of AUC to 0.71.

Even when the PAF is 50%, the AUC remains less than 0.64 for the ranges of genotype frequency and relative risks considered. When PAF is 50%, genotype frequency 0.1 and relative risk 1.1, the number of markers that contribute to risk of disease is 70 (AUC=0.57). When we increase the relative risk to 1.5, the number of markers decline to 14 (AUC=0.64). This is the maximum AUC obtained when PAF is 50%. Increasing the genotype frequency to 0.3 results in the number of markers declining further to 5 (AUC=0.62). These results show that for a given value of PAF, the number of markers that contribute to risk of disease declines with increasing genotype frequency and results in lower AUC; however, the number of markers that contribute to risk of disease also declines with increasing relative risks but leads to higher

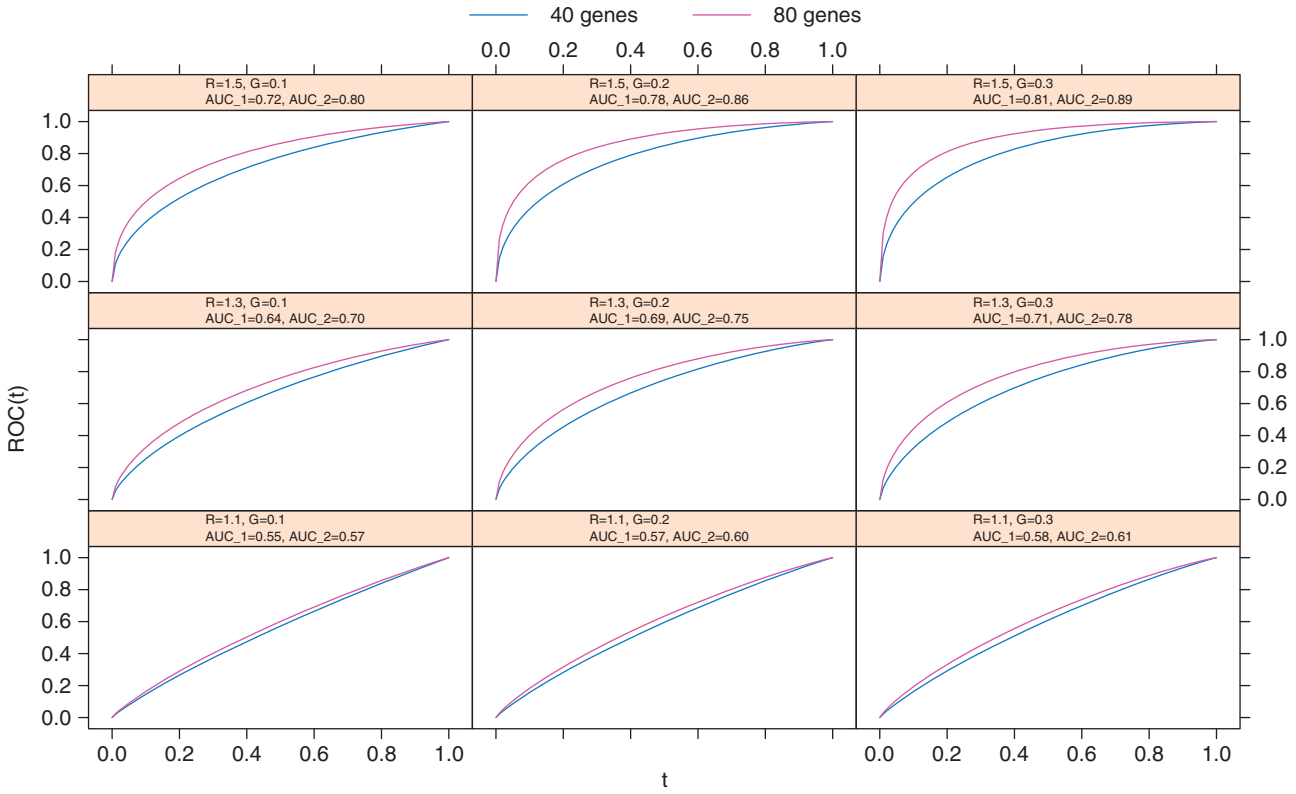

Figure 1 Area under the ROC curve (AUC) for genotype frequency (*G*) equal to 0.1, 0.2 and 0.3, and relative risks (*R*) equal to 1.1, 1.3 and 1.5, for 40 markers (AUC_1) and 80 markers (AUC_2).

AUC. When the relative risk is 1.5, the PAF has to be at least 84% to have a reasonable discriminative power (AUC around 0.70) for the range of genotype frequency considered. When the relative risk is 1.1 and genotype frequency 0.3, even with a PAF of 99% the AUC is only 0.65.

When genotype frequency and relative risks differ for different markers, we used the formula derived for the best linear combination of markers to calculate the AUC. We also simulated one million observations for a case control study for specified genotype frequencies for cases and controls for multiple markers assuming a disease prevalence of 10%. The logistic procedure in SAS (Cary, NC, USA) was used to calculate the concordance statistic for the simulated data. Tables 1a–c give the AUC calculated using the best linear combination of markers (AUCB) and the AUC calculated from the logistic procedure (AUCS) for different values of relative risks and different genotype frequencies for two, three and five markers, respectively. For two or three markers, AUCB is always greater than AUCS. This is to be expected because the AUCS is based on the empirical ROC curve and the AUCB is based on the binormal ROC curve. The maximum difference between AUCS and AUCB was less than 0.024 for all the ranges of genotype frequencies (0.1–0.4) and relative risks (1.1–2.0) considered. For five markers, there is almost no difference between AUCB and AUCS. These tables show again that AUC increases with increasing genotype frequency, increasing relative risks and increasing number of markers in the genomic profile.

Example

Zheng *et al.*[18] studied the genetic predisposition to prostate cancer by examining the association between prostate cancer and five SNPs that map to the three 8q24 loci, to 17q12 and to 17q24.3. Individually, the risk ratios associated with these loci ranged from 1.22 to 1.53. Table 2 gives the odds ratios adjusted for age and geographic region, and genotype frequency of the five SNPs for prostate cancer susceptibility (AUC=0.58).

Let the genotype frequencies for the five SNPs are given by G_1=0.30, G_2=0.25, G_3=0.07, G_4=0.77 and G_5=0.26, and the relative risks by R_1=1.38, R_2=1.28, R_3 =1.53, R_4 =1.37 and R_5=1.22. The AUC calculated using the formula (AUCB) is 0.58. The joint PAF using the formula given in Zheng *et al.*[18] for the five SNPs is 0.4045. When we consider five identical markers with average relative risk 1.36 and average genotype frequency 0.33, and PAF 0.40, the AUC calculated using the formula for PAF is 0.59.

DISCUSSION

We provide a direct method to evaluate the clinical discriminative accuracy of a set of polymorphisms in a genomic profile when genotype frequency and relative risks of each polymorphism are known. The comparison of AUCs obtained from our method and the simulation using logistic regression show that our method provides almost identical AUC. There have been some concerns of using the Fisher's LDF for binary data. For example, for two genetic variants, the Fisher's LDF will behave poorly for those situations where either both genetic variants are present or none of them are present occur more frequently in one population (cases or controls) whereas only one of the two genetic variants is present occur more frequently in the other population.[19] These situations never occurred in our simulation study for the ranges of genotype frequency and relative risks considered in this paper.

Our results show that both genotype frequencies and relative risks are equally important factors for predicting common diseases.

Table 1a Areas under the ROC curves for the best linear combination of markers (AUCB) and for simulated data using logistic regression (AUCS) for two markers

G1	*G2*	*R1*	*R2*	*AUCS*	*AUCB*
0.1	0.1	1.1	1.1	0.508	0.512
0.1	0.1	1.1	1.5	0.524	0.538
0.1	0.1	1.5	1.5	0.538	0.552
0.1	0.1	2	2	0.571	0.594
0.1	0.4	1.1	2	0.588	0.597
0.1	0.4	2	1.1	0.549	0.568
0.4	0.1	1.1	1.5	0.530	0.539
0.4	0.1	1.5	1.1	0.552	0.557
0.4	0.4	1.1	1.1	0.517	0.519
0.4	0.4	1.5	1.1	0.555	0.558
0.4	0.4	1.5	1.5	0.575	0.580
0.4	0.4	2	2	0.626	0.636

Table 1b Areas under the ROC curves for the best linear combination of markers (AUCB) and for simulated data using logistic regression (AUCS) for three markers

G1	*G2*	*G3*	*R1*	*R2*	*R3*	*AUCS*	*AUCB*
0.1	0.1	0.1	1.1	1.1	1.1	0.511	0.514
0.1	0.1	0.1	1.1	1.5	2.0	0.559	0.576
0.1	0.1	0.1	1.1	2.0	2.0	0.574	0.594
0.1	0.1	0.1	2.0	2.0	2.0	0.596	0.614
0.1	0.1	0.4	1.1	1.1	1.5	0.554	0.558
0.1	0.4	0.1	1.1	1.1	1.5	0.531	0.540
0.1	0.2	0.4	1.1	1.5	2.0	0.605	0.608
0.1	0.2	0.4	2.0	2.0	2.0	0.633	0.643
0.4	0.2	0.1	1.1	1.5	2.0	0.573	0.583
0.4	0.4	0.4	1.1	1.1	1.1	0.521	0.523
0.4	0.4	0.4	1.1	1.1	1.5	0.558	0.560
0.4	0.4	0.4	1.1	1.5	1.5	0.579	0.581
0.4	0.4	0.4	1.5	1.5	1.5	0.593	0.600
0.4	0.4	0.4	2.0	2.0	2.0	0.657	0.665

Table 1c Areas under the ROC curves for the best linear combination of markers (AUCB) and for simulated data using logistic regression (AUCS) for five markers

G1	*G2*	*G3*	*G4*	*G5*	*R1*	*R2*	*R3*	*R4*	*R5*	*AUCS*	*AUCB*
0.1	0.1	0.1	0.1	0.1	1.1	1.1	1.1	1.1	1.1	0.516	0.518
0.1	0.1	0.1	0.1	0.1	1.1	1.2	1.4	1.8	2.0	0.581	0.593
0.2	0.2	0.2	0.2	0.2	1.1	1.2	1.4	1.8	2.0	0.617	0.619
0.1	0.2	0.3	0.3	0.4	1.1	1.2	1.4	1.8	2.0	0.634	0.634
0.2	0.2	0.2	0.3	0.3	1.2	1.2	1.2	1.3	1.3	0.559	0.561
0.2	0.2	0.2	0.3	0.3	1.2	1.2	1.2	1.3	2.0	0.605	0.606
0.3	0.3	0.2	0.2	0.1	1.1	1.2	1.4	1.8	2.0	0.604	0.608
0.2	0.2	0.3	0.3	0.4	1.1	1.2	1.4	1.8	2.0	0.634	0.634
0.2	0.2	0.3	0.4	0.4	1.1	1.2	1.4	1.8	2.0	0.635	0.636
0.3	0.3	0.3	0.3	0.3	1.1	1.2	1.4	1.8	2.0	0.633	0.633
0.4	0.3	0.3	0.2	0.1	1.1	1.2	1.4	1.8	2.0	0.608	0.610
0.2	0.2	0.3	0.3	0.4	1.2	1.2	1.2	1.2	1.2	0.549	0.552
0.4	0.4	0.4	0.4	0.4	1.1	1.2	1.4	1.8	2.0	0.637	0.637
0.4	0.4	0.4	0.4	0.4	2.0	2.0	2.0	2.0	2.0	0.704	0.709

Table 2 Genotype frequency and odds ratios of five SNPs for prostate cancer susceptibility

SNP	*Chromosomal region*	*Risk group*	*Genotype frequency (G)*	*Odds ratio (R)*	*PAF*
rs4430796	17q12	CC/TC *vs* TT	0.30	1.38	0.1023
rs1859962	17q24.3	GT/TT *vs* GG	0.25	1.28	0.0654
rs16901979	8q24 (region 2)	CC *vs* AA/CA	0.07	1.53	0.0358
rs6983267	8q24 (region 3)	TT *vs* GT/GG	0.77	1.37	0.2217
rs1447295	8q24 (region 1)	CC *vs* CA/AA	0.26	1.22	0.0541

The discriminative accuracy increases with increasing genotype frequency, increasing relative risk and increasing number of risk genotypes. If lower bounds and upper bounds of relative risks and genotype frequency are available, one can use the formula for AUC given in this paper to obtain the empirical distribution of AUC by considering a range of values of genotype frequency and relative risks between these two bounds.

For a given value of PAF, the discriminative accuracy increases with increasing relative risks, but declines with increasing genotype frequency. The joint PAF for the five SNPs given in the example is 0.40 and the AUC is 0.59. The clinical discriminative accuracy of these five SNPs in a genomic profile for prostate cancer is very low and the planned marketing of a genetic test based on this study[20] is clearly premature. Assuming the average relative risk (1.36) and average genotype frequency (0.33) in this example for each risk genotype, 25 risk genotypes are required to be in the genomic profile (PAF=0.94) to have a reasonable clinical discriminative accuracy for prostate cancer.

For given values of relative risk and genotype frequency, the discriminative accuracy increases with increasing heritability but declines with increasing prevalence of the disease. An analysis of monozygotic and dizygotic twin pairs in Scandinavia concluded that 42% (95% confidence interval (29–50%)) of prostate cancer risk may be accounted for by heritable factors.[21] The prevalence of diagnosed prostate cancer in US adult population is about 1.6% based on the estimates from the National Health Interview Survey. Assuming an average relative risk of 1.36 and a genotype frequency of 0.33, the clinical discriminative accuracy is 90% for 146 risk genotypes. Because the true prevalence of prostate cancer is unknown, assuming an upper bound of 3.2% for the prevalence of prostate cancer, the clinical discriminative accuracy declines to 87% for 115 risk genotypes.

Our study is limited by the assumption of independence of the genetic variants, and our inability to model gene–gene interactions. We also assumed multiplicative risk models. Joint genetic effects on risk may be neither multiplicative nor additive. Unfortunately, for statistical modeling, epidemiological analyses have had to deal with multiplicative or additive models. Multiplicative risk models are expected to yield higher predictive accuracy than that of additive models because the joint effect of risks for multiplicative models is higher than the joint effect of risks for additive models.

If the genetic variants are in linkage disequilibrium, the variance covariance matrix of the exposure variables will contain nonzero covariance terms for both cases and controls. Assuming binary environmental risk factors and the independence of genes and environmental risk factors in controls, gene–environmental interactions would lead to nonzero covariance terms in the variance covariance matrix of the exposure variables in cases. It is known that the LDF does not perform well when binary variables are correlated. More methodological work is needed in this area to

study the proper discriminant function for the range of genotype frequency and relative risks considered in this paper.

Although our calculations assume an unrealistic scenario of identical risk genotypes, we clearly show that to have an improved risk prediction for any common disease, many more risk genotypes associated with a disease are required in a genomic profile than the ones currently available. It is clear that the five SNPs in the example do not explain the full familial aggregation for prostate cancer supporting the existence of additional loci for prostate cancer. A study of 160 unique polymorphisms-disease associations included in commercial genomic profiles found only 29 polymorphisms significantly associated with the diseases in meta-analyses.[22] There were 33 diseases significantly associated with one or more of these 29 polymorphisms and the maximum number of polymorphisms significantly associated with a given disease was only 3. Until a sufficient number of polymorphisms significantly associated with a disease are found that would provide a reasonable clinical discriminative accuracy, genomic profiling is currently not an effective tool for clinical applications.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

1 Morris JK, Wald NJ: Graphical presentation of distributions of risk in screening. *J Med Screen* 2005; **12**: 155–160.
2 Zhen Z, Yu Y, Berchuck A *et al*: Combining multiple serum tumor markers improves detection of stage I epithelial ovarian cancer. *Gynecol Oncol* 2007; **107**: 526–531.
3 Wald NJ, Morris JK, Rish S: The efficacy of combining several risk factors as a screening test. *J Med Screen* 2005; **12**: 197–201.
4 Ioannidis JP: Genetic associations: false or true? *Trends Mol Med* 2003; **9**: 135–138.
5 Khoury MJ, Yang Q, Gwinn M, Little J, Flanders WD: An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions. *Genet Med* 2004; **6**: 38–47.
6 Hanley JA: Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diag Imaging* 1989; **29**: 307–335.
7 Pepe MS: Receiver operating characteristic methodology. *J Am Stat Assoc* 2000; **95**: 308–311.
8 Zhou XH, McClish DK, Obuchowsk NA: *Statistical Methods in Diagnostic Medicine*. New York, USA: John Wiley & Sons, 2002.
9 Janssens ACJW, Moonesinghe R, Yang Q, Steyerberg EW, Duijn CM, Khoury MJ: The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic disease. *Genet Med* 2007; **9**: 528–535.
10 Moonesinghe R, Yang Q, Khoury MJ: Sample size required to detect a system of genetic variants. *Emerg Themes Epidemiol* 2008; **5**: 24.
11 Wray NR, Goddard ME, Visscher PM: Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res* 2007; **17**: 1520–1528.
12 Moonesinghe R: A refinement to 'how many genes underlie the occurrence of common complex diseases in the population? *Int J Epidemiol* 2006; **35**: 497.
13 Su JQ, Liu JS: Linear combinations of multiple diagnostic markers. *J Am Stat Assoc* 1993; **88**: 1350–1355.
14 Moore DH: Evaluation of five discriminant procedures for binary variables. *J Am Stat Assoc* 1973; **68**: 399–404.
15 Hand DJ: A comparison of two methods of discriminant analysis applied to binary data. *Biometrics* 1983; **39**: 683–694.
16 Ganeshanandam S, Krzanowski WJ: Error-rate estimation in two-group discriminant analysis using the linear discriminant function. *J Statist Comput Simulation* 1990; **36**: 157–175.
17 Asparoukhov OK, Krzanowski WJ: A comparison of discriminant procedures for binary variables. *Comput Stat Data Anal* 2001; **38**: 139–160.
18 Zheng SL, Sun MDJ, Wiklund F *et al*: Cumulative association of five genetic variants with prostate cancer. *N Engl J Med* 2008; **358**: 910–919.
19 Krzanowski WJ: The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics* 1977; **19**: 191–200.
20 Kolata G: *$300 to Learn Risk of Cancer of the Prostate*. New York, USA: The New York Times Company, 2008.
21 Lichtenstein P, Holm NV, Verkasalo PK *et al*: Environmental and heritable factors in the causation of cancer – Analysis of cohorts of twins from Sweden, Denmark and Finland. *N Engl J Med* 2000; **343**: 78–85.
22 Janssens ACJW, Gwinn M, Bradley LA, Oostra BA, Cornellia M, Dujin CM: A Critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions. *Am J Hum Genet* 2008; **82**: 593–599.

APPENDIX A

Let P be the prevalence of disease, h^2 the heritability of the disease on the observed scale, G the genotype frequency, R the relative risk and k the number of loci.

Let d=Prob (affected| genotype)=IR^X, where I is the background risk and X is the number of genotypes. Then,

$$P = E(d) = \sum_{X=0}^{k} IR^X G^X (1-G)^{(1-X)} = I[RG+(1-G)]^k$$

$$\begin{aligned} V(d) = E(d^2) - [E(d)]^2 &= I^2 \sum_{X=0}^{k} IR^{2X} G^X (1-G)^{(1-X)} - I^2[RG+(1-G)]^{2k} \\ &= I^2[R^2G+(1-G)]^k - I^2[RG+(1-G)]^{2k} \\ &= \frac{P^2[R^2G+(1-G)]^k}{[RG+(1-G)]^{2k}} - P^2 \end{aligned} \qquad (A.1)$$

The variance of disease prevalence due to genetic factors is $h^2\,P(1-P)$. Therefore,

$$h^2P(1-P) = \frac{P^2[R^2G+(1-G)]^k}{[RG+(1-G)]^{2k}} - P^2$$

and

$$\frac{h^2(1-P)+P}{P} = \frac{[R^2G+(1-G)]^k}{[RG+(1-G)]^{2k}}$$

The number of loci k is given by

$$k = [\log\{h^2(1-p)+p\} - \log p]/[\log\{R^2G+(1-G)\} - 2\log\{RG+(1-G)\}]$$