ARTICLE

Using the Optimal Receiver Operating Characteristic Curve to Design a Predictive Genetic Test, Exemplified with Type 2 Diabetes

Qing Lu[1] and Robert C. Elston[1,*]

Current extensive genetic research into common complex diseases, especially with the completion of genome-wide association studies, is bringing to light many novel genetic risk loci. These new discoveries, along with previously known genetic risk variants, offer an important opportunity for researchers to improve health care. We describe a method of quick evaluation of these new findings for potential clinical practice by designing a new predictive genetic test, estimating its classification accuracy, and determining the sample size required for the verification of this accuracy. The proposed predictive test is asymptotically more powerful than tests built on any other existing method and can be extended to scenarios where loci are linked or interact. We illustrate the approach for the case of type 2 diabetes. We incorporate recently discovered risk factors into the proposed test and find a potentially better predictive genetic test. The area under the receiver operating characteristic (ROC) curve (AUC) of the proposed test is estimated to be higher (AUC = 0.671) than for the existing test (AUC = 0.580).

Introduction

With the latest improvements and ever-decreasing costs of high-throughput genotyping technologies, large-scale genetic-association studies, and in particular genome-wide association studies, are now being conducted. These studies provide a comprehensive scan of the whole genome and have the potential to identify many more genetic risk variants for common complex diseases. The new findings from these studies, together with previously known genetic and environmental risk factors—whether or not they increase our understanding of the etiology of common complex diseases—offer a potential opportunity for researchers to improve medical care and public health.[1]

Some previous efforts aimed at combining genetic and environmental findings to predict disease or, more precisely, to develop a predictive genetic test, are discussed elsewhere.[2–5] These are important first steps toward the development of successful predictive genetic tests for common complex diseases,[6] and such tests have been recognized as comprising the cornerstone of future genomic medicine.[7] The hope is that these tests will provide an early discovery of an individual's disease risk so that appropriate prevention strategies can be used to reduce morbidity and mortality. These tests are anticipated to have a large impact on health care and change the form of health care away from treatment toward prevention.[8]

However, these previous efforts have been limited by the genetic and clinical information available at the time the predictive genetic tests were developed. For instance, the current predictive genetic test for type 2 diabetes is based on three common variants.[3] With the genome-wide association studies that have now been conducted for this disease and the novel susceptibility variants identified,[9,10] one may want to know how much an existing test could be improved by incorporating into it the newly discovered genetic susceptibility variants.

The identification of risk variants is a progressive process, and so predictive genetic tests will also be subject to change whenever novel genetic susceptibility variants are discovered. This requires a flexible and easily implemented method for the redevelopment of predictive genetic tests with time.[6] In this paper, we propose a flexible model to help design a new predictive genetic test. Using information garnered from published genetic-association studies, clinical studies, and even a previously accepted predictive genetic test, this approach provides an estimated classification accuracy of the proposed test and hence an idea of how much improvement over an existing test a proposed new test might achieve.

The clinical performance and applicability of a predictive genetic test rests on four main components: (1) analytic validity, (2) clinical validity, (3) clinical utility, and (4) associated ethical, legal, and social implications.[11] Our work here focuses on addressing the clinical validity issue, defined as the ability of a genetic test to detect or predict the associated disorder (phenotype). The assessment of clinical validity is an important step and is the starting point for test building. If the predictive genetic test discriminates well among possible eventual outcomes, we continue to evaluate its clinical utility. If, on the other hand, it has poor accuracy, it is unlikely to have practical value for patient care.[12]

In this paper, we first derive a general formula to build a predictive genetic test from previous independent association results, extending it to situations where the genetic variants are in linkage disequilibrium or interact with each other. We then indicate how the requisite sample

[1]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106, USA
*Correspondence: rce@darwin.case.edu
DOI 10.1016/j.ajhg.2007.12.025.

size can be calculated in order to design a study that will have specified power at a given significance level to be sure that this new test attains a desired level of accuracy or has a higher level of accuracy than the existing test. Finally, we illustrate the model with recent association results that have become available for type 2 diabetes.

Material and Methods

Several approaches have been proposed for the evaluation of a predictive genetic test based on multiple disease-susceptibility loci. Among them, the receiver operating characteristic (ROC) curve[13] has been recognized as the most suitable measure.[14,15] The ROC curve plots a test's true-positive rate (sensitivity) against its false-positive rate (1-specificity) for continuously changing cutoffs over the whole possible range of test results. It evaluates the tests that result from these cutoffs with the entire spectrum of pairs of true-positive rates (TPRs) and false-positive rates (FPRs) and so gives a global description of a test's classification accuracy.[16] The ROC curve is also one of the most popular measures used for clinical diagnostic tests and has been widely used in different areas of medicine.[17]

Among all the techniques available for combining multiple predictors for ROC analysis, the logistic-regression-based approach has been the most commonly used. However, as recent studies have shown, logistic regression will not be optimal if the logistic model does not hold.[18] In the statistics literature, the optimality of the likelihood ratio for ROC analysis has long been recognized.[19,20] The likelihood ratio—which in this context is defined as the ratio of two density functions of a predictor x, that among cases to that among controls—is useful for the generation of the appropriate (optimal) ROC curve. With Egan's definition,[20] the optimal ROC curve results when this likelihood ratio is plotted from its largest value to its smallest value. The optimal ROC curve displays the best possible performance set based on the likelihood ratio in terms of (1) the maximization of the TPR for any fixed value of the FPR, (2) the minimization of the overall misclassification probability, and (3) the minimization of the expected cost. These three ideal properties could be obtained simply by the application of the Neyman-Pearson lemma.[20,21] Although the likelihood ratio is ideal for ROC analysis, it was only recently emphasized for combining multiple predictors.[21,22] Baker,[22] who noted the important role of the likelihood ratio for combining multiple predictors, based his argument regarding its optimality by drawing on cost-effectiveness theory[23] rather than on the Neyman-Pearson lemma. Despite this difference, the two arguments are essentially the same, suggesting that for multiple predictors, a test based on the likelihood ratio is optimal for any fixed value of TPR or FPR. This conclusion is elegant, but it has limitations if we are interested in comparing tests with different TPR and FPR pairs or in evaluating the overall performance of the test. For the latter purpose, the ROC curve or its summary indexes are the appropriate criteria to use. Here we choose the area under the ROC curve (AUC), the most popular summary index of the ROC curve, as a measure of a test's performance and prove that the optimal ROC curve has the highest AUC (Appendix A). The AUC measures the probability that test values from a randomly selected pair of diseased and nondiseased individuals are correctly ordered and is thus a convenient global measure for the quantification of classification (diagnostic) accuracy.[12] This means that a test based on the optimal ROC curve achieves the highest classification accuracy among all approaches (including a logistic-regression-based approach).

Building a Predictive Genetic Test on the Basis of the Optimal ROC Curve

A General Model for Independent, Noninteracting Loci

Suppose we are interested in constructing a predictive genetic test on the basis of variants at n genetic loci. From previous association studies or other public sources, we obtain the relative risks estimates ($R_i = (r_{i1},\cdots,r_{im_i})$) and population genotype frequency estimates ($F_i = (f_{i1},\cdots,f_{im_i})$) for the ith ($i = 1,\cdots,n$) associated locus, which has m_i possible genotypes. The distribution of these genotypes at the ith locus in the disease (D) population can then be derived from this information, with the formula

$$P(g_{ij_i}|D) = \frac{r_{ij_i} f_{ij_i}}{\sum_{j_i=1}^{m_i} r_{ij_i} f_{ij_i}}, j_i = 1,\cdots,m_i. \quad (1)$$

The m_i equations displayed in Equation 1 are appropriate for a variety of genetic variants (in particular, single-nucleotide polymorphisms) with known mode of inheritance (Appendix B). The equations can also be modified for the situation where the risk estimates are odds ratios (Appendix B). The information we can use is not limited to that from association studies; it can also come from other genetic studies (Appendix B).

If we assume linkage equilibrium among the n loci, we can calculate the joint probability of the multilocus genotype $G_k = (g_{1j_1},g_{2j_2},\cdots,g_{nj_n})$ from the single-locus-genotype frequencies,

$$P(G_k) = \prod_{i=1}^{n} P(g_{ij_i}) = \prod_{i=1}^{n} f_{ij_i}, j_i = 1,\cdots,m_i, k = 1,\cdots,K, \quad (2)$$

where K denotes the total number of multilocus genotypes possible from the n disease-susceptibility loci and its maximal value is $m_1 \cdot m_2 \cdot \ldots \cdot m_n$. If we further adopt a multiplicative model, which assumes that the joint effect of the n genetic variants is proportional to the product of the individual variants' main effects, then the probabilities of the multilocus genotypes G_k given disease status can be expressed as

$$\begin{cases} P(G_k\,|\,D) = \prod_{i=1}^{n} P(g_{ij_i}\,|\,D) \\ P(G_k\,|\,\overline{D}) = \dfrac{P(G_k) - P(G_k\,|\,D)\rho}{1-\rho} \end{cases}, j_i = 1,\cdots,m_i, k = 1,\cdots,K, \quad (3)$$

where $\overline{D}$ denotes absence of disease and ρ denotes the disease prevalence. Given these probabilities, we can calculate the likelihood ratios (LRs):

$$LR_k = \frac{P(G_k\,|\,D)}{P(G_k\,|\,\overline{D})}, k = 1,\cdots,K. \quad (4)$$

We rank the multilocus genotypes in descending order of their LRs, from the highest rank to the lowest rank, and plot the test's TPR (sensitivity) against its FPR (1-specificity) for each possible cutoff between adjoining pairs of multilocus genotypes that might be used in the prediction of disease. This gives us the empirical optimal ROC curve, which simply consists of a set of TPR and FPR pairs:

$$\begin{cases} TPR_{(k)} = \sum_{\kappa=1}^{k} P\big(G_{(\kappa)}|D\big) \\ FPR_{(k)} = \sum_{\kappa=1}^{k} P\big(G_{(\kappa)}|\overline{D}\big) \end{cases}, \; k = 1,\cdots,K, \quad (5)$$

where $G_{(\kappa)}$ is the κ th genotype in the sequence of likelihood ratios. Because the LRs are in descending order and the LRs correspond to the slope of the ROC curve, this indicates that the optimal ROC curve is always concave.

Once the optimal ROC curve has been built, we obtain the explicit expression for the area under the optimal ROC curve by applying the trapezoid rule:

$$AUC = \frac{1}{2}\sum_{k=1}^{K}\left(TPR_{(k)} + TPR_{(k-1)}\right)\cdot\left(FPR_{(k)} - FPR_{(k-1)}\right), \qquad (6)$$

where $TPR_{(0)} = FPR_{(0)} = 0$. This measures the estimated discriminative ability of the test and leads to the highest value of the AUC among all approaches to designing a predictive test. Other statistics we might be interested in, such as predictive values, can be directly obtained from the optimal ROC curve (Appendix C). Proof of the optimality of the predictive values obtained this way follows from the fact that the optimal ROC curve maximizes the TPR for any fixed value of the FPR and the equations in Appendix C.

Genetic Loci in Linkage Disequilibrium or Interacting with Each Other

The above model can be extended to incorporate loci that are in linkage disequilibrium (LD) with each other. In this case, we are interested in the multilocus genotypes formed by these linked loci. Assume we have L linked loci, and for each locus we have $K_l\,(l = 1,\cdots,L)$ alleles. Following a notation similar to that in Gorelick and Laubichler,[24] we denote by $D_n(k_l,\cdots)$ the coefficients of LD between n loci $(n = 2,\cdots,L)$ and by $D_1(k_l)$ the population allele frequency for the k_l th $(k_l = 1,\cdots,K_l)$ allele at locus l. Assuming that all possible LD coefficient and population allele-frequency estimates can be obtained from previous studies, we can express the haplotype frequency (p_h) for haplotype $h = (k_1, k_2,\ldots, k_L)$ as the summation of all possible products of LD coefficients and allele frequencies whose orders (i.e., the number of loci) add to L,

$$p_h = \sum_{\sum_{s=1}^{S} n_s = L}\left[\prod_{s=1}^{S} D_{n_s}(\cdots)\right], \qquad (7)$$

where D_{n_s} is the coefficient of linkage disequilibrium between n_s loci[24].

Normally, the coefficients of LD with order higher than two are rarely reported in genetic studies. Denote by $D_2(k_l, k_{l'})$ the pairwise LD between the alleles $k_l\,(k_l = 1,\cdots,K_l)$ and $k_{l'}\,(k_{l'} = 1,\cdots,K_{l'})$ at loci l and $l'\,(l < l'; l = 1,\cdots,L-1)$. Assuming that for each pair of loci l and l' all possible $(K_l - 1)\times(K_{l'} - 1)$ pairwise LD coefficient estimates can be obtained from previous studies (in particular, $(K_l - 1)\times(K_{l'} - 1) = 1$ for SNPs), we can approximate the haplotype frequency by only using the pairwise LD (D_2) and the population allele frequencies (D_1),

$$p_h \approx \sum_{\substack{\sum_{s=1}^{S} n_s = L \\ n_s \le 2}}\left[\prod_{s=1}^{S} D_{n_s}(\cdots)\right]. \qquad (8)$$

Given the haplotype frequency and the assumption of Hardy-Weinberg equilibrium (HWE), we can derive the distribution for the phased multilocus genotype $g_j = (hh')$,

$$f_j = p(g_j) = \begin{cases} 2p_h p_{h'} & h \neq h' \\ p_h^2 & h = h' \end{cases},\ j = 1,\cdots,m.$$

Provided that the haplotype-relative risk information r_h is available, we can derive the relative risks for the phased multilocus genotype $g_j = (hh')$, on the basis of an additive model, as

$$r_{g_j} = r_h + r_{h'}, j = 1,\cdots,m.$$

Although the above equations assume an additive model, any other model (e.g., recessive) can also be adopted according to any prior knowledge of the disease. By treating these L linked loci as comprising one set of genotypes and applying Equation 1, we can incorporate linked loci into the approach.

In a similar manner, we can extend the model to handle interacting loci. In this scenario, we group all possible multilocus genotypes from the interacting loci into a few clusters, each with a different associated disease risk. In the simplest situation, we have just two clusters, a high-risk cluster and a low-risk cluster. At the other extreme, each multilocus genotype itself represents a cluster. Then, by obtaining the relative risks and the distribution of these clusters, and again applying Equation 1, we can incorporate interacting loci into the model.

We illustrate this by using a simple example. Assume there is an interaction between two SNPs (A and B) and the underlying interaction follows a threshold model,[25] defined as implying there is a single high risk for all individuals having at least one of the disease-susceptibility alleles at each of the two loci and a common low risk for all other individuals. We denote by rr the relative risk of the high-risk group (g_1) compared to the low-risk group (g_0) and by f the population frequency of the high-risk group. Then the distribution of the high- and low-risk groups in cases can be written as

$$\begin{cases} P(g_1 \mid D) = \dfrac{rr\cdot f}{rr\cdot f + 1 - f} \\ P(g_0 \mid D) = \dfrac{1 - f}{rr\cdot f + 1 - f_0} \end{cases},$$

where, assuming Hardy-Weinberg equilibrium and that p_A and p_B are the frequencies of the disease-susceptibility alleles for the two loci, $f = p_A(2 - p_A)p_B(2 - p_B)$. Although we illustrate the model by using genetic variants, the equations also apply for clinical risk factors and to the situation where there is gene-environment interaction.

Sampling Variability of the Empirical AUC and Comparison of Empirical ROC Curves

The above procedure provides an estimated classification accuracy of the new test in terms of the AUC. The estimated AUC could be subject to large variability if the sample size is small. We introduce here two methods for the calculation of the AUC variance, which enables us to quantify its precision.

Asymptotically, the variance of the estimated AUC depends only on the ROC curve itself and the numbers of cases and controls:[26]

$$\mathrm{var}_A(AUC) = \frac{\mathrm{var}_D}{n_{\overline{D}}} + \frac{\mathrm{var}_{\overline{D}}}{n_D}, \qquad (9)$$

where n_D and $n_{\overline{D}}$ are the sample sizes of the disease and nondisease samples and var_D and $\mathrm{var}_{\overline{D}}$ are given by

$$\begin{cases} \mathrm{var}_D = \left[\frac{1}{2}\sum_{k=1}^{K}\left(TPR_{(k)} + TPR_{(k-1)}\right)^2\cdot\left(FPR_{(k)} - FPR_{(k-1)}\right)\right] - AUC^2 \\ \mathrm{var}_{\overline{D}} = \left[\frac{1}{2}\sum_{k=1}^{K}\left(FPR_{(k)} + FPR_{(k-1)}\right)^2\cdot\left(TPR_{(k)} - TPR_{(k-1)}\right)\right] - (1 - AUC)^2 \end{cases}. \qquad (10)$$

This provides a simple variance estimate when all the associated genetic variants used for the test are based on one study. If the genetic variants come from different studies, we cannot use the above equations because n_D and $n_{\overline{D}}$ are not defined. In this case, we can adopt a bootstrap approach to estimate the variance. Suppose the n genetic variants for the test are from U independent studies. For the ith ($i = 1,\cdots, n_u$) associated variant in the uth study ($u = 1,\cdots, U$), we apply Equation 1 to compute its genotype frequency in the case sample and repeat this step for all n_u genetic variants in the uth study. Further, by using Equations 2 and 3 to combine these n_u genetic variants, we can obtain the probabilities of the multilocus genotypes given disease status. Given the total number of diseased and nondiseased individuals in the uth study, we can derive the observed numbers of all possible multilocus genotypes given disease status. On the basis of these observed numbers, we draw a bootstrap sample and use the sample to calculate the genotype frequencies, given disease status, for each of the n_u variants. We repeat this procedure for all the other U-1 studies and then apply Equations 2–6 to construct the optimal ROC curve and compute the AUC estimate. By drawing a large number of bootstrap samples (e.g., 1000), we can obtain the bootstrap variance for the AUC estimate, denoted $\mathrm{var}_B(AUC)$. Although we illustrate the bootstrap approach for genetic variants, the same approach also applies for clinical risk factors and to the situation where the genetic loci are in LD or interacting with each other.

With the variance estimates, we can easily determine the significance of the difference between two AUC estimates, A_1 and A_2, for two different predictive tests. If the associated variants on which the two tests are based come from different studies, the variance of the AUC difference is equivalent to the sum of the variances of the two AUC estimates,

$$\mathrm{var}(A_1 - A_2) = \mathrm{var}(A_1) + \mathrm{var}(A_2)\,,$$

where $\mathrm{var}(A_1)$ and $\mathrm{var}(A_2)$ are the variances of A_1 and A_2. On that basis, we construct an appropriate test statistic that under the null hypothesis and in large samples follows a standard normal distribution:

$$\frac{A_1 - A_2}{\sqrt{\mathrm{var}(A_1 - A_2)}}. \qquad (11)$$

If some or all of the associated variants on which the two tests are based come from the same studies, then we need to take the covariance of A_1 and A_2 into account. We can adopt the same bootstrap approach here, but now forming two optimal ROC curves from each bootstrap sample. The bootstrap variance of the AUC difference can be expressed as

$$\mathrm{var}_B(A_1 - A_2) = \mathrm{var}_B(A_1) + \mathrm{var}_B(A_2) - 2\mathrm{cov}_B(A_1, A_2)\,,$$

where $\mathrm{var}_B(A_1)$, $\mathrm{var}_B(A_2)$, and $\mathrm{cov}_B(A_1, A_2)$ are the bootstrap variances of A_1 and A_2 and the bootstrap covariance between A_1 and A_2.

Sample-Size Calculation

If the classification accuracy estimate (i.e., the AUC) of the new test appears to be superior to existing tests, or if it reaches a desired accuracy level, it might be worth further developing for clinical use. However, the clinical validity of the test, i.e., its classification and/or prediction accuracy, should be comprehensively evaluated before considering the test for clinical use.[11] For that purpose, a new replication study is necessary.[27] Such a study serves the purpose of verifying the test's estimated classification accuracy, which has been so far estimated on the basis of assumptions and information from published genetic studies. To conduct the study, we set up a hypothesis of interest and design the study with the requisite sample size to test that hypothesis. If we assume A_0 is the AUC that measures the classification accuracy of a previous test, or is the minimum desired level of classification accuracy, we are interested in knowing whether the performance of the new test is superior to A_0. Our null hypothesis for this purpose is $H_0 : A = A_0$, with the alternative $H_A : A > A_0$, where A is the AUC for the new test. An appropriate sample size can then be determined that will ensure, with specified power at an appropriate significance level, that the new test exceeds the minimal acceptable AUC value A_0. For this, we adopt the general approach based on asymptotic theory.[28] Assume α and $1 - \beta$ are the specified type I error and power we require for our test; the required sample size for the test can then be expressed as

$$n_D = (p\,\mathrm{var}_D + \mathrm{var}_{\overline{D}})\left\{\frac{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)}{A - A_0}\right\}^2, \qquad (12)$$

where n_D and $n_{\overline{D}}$ are the sample sizes required for the disease and nondisease samples, respectively, $p = n_D/n_{\overline{D}}$, and Φ is the standard normal cumulative distribution function.

Application to Type 2 Diabetes

With the numerous genetic and clinical studies conducted so far, our understanding of the causes of type 2 diabetes has greatly improved. Thus, now might be the right time to begin applying the more recent findings into clinical use, developing for type 2 diabetes a predictive test that combines all possible genetic variants and environmental factors. In particular, we use here novel susceptibility loci that have been identified in recent genome-wide association studies conducted for type 2 diabetes.[9,10]

Searching for a successful predictive genetic test for type 2 diabetes has already been initiated. Recently, Weedon et al.[3] used three common variants, rs5219 (Glu23Lys) of KCNJ11, rs1801282 (Pro12Ala) of PPARG, and rs7903146 of TCF7L2, to predict the risk of type 2 diabetes. We start with their study to test how consistent the result from our approach is with their findings and then investigate how much we might be able to improve their test by utilizing the newer findings from genome-wide studies.

From the study by Weedon et al.,[3] we obtained the allele frequencies among cases and controls for the three variants. We computed the genotype frequencies among cases and controls assuming Hardy-Weinberg equilibrium for both the case and the control populations. Applying Equation 3, we computed the multilocus genotype probabilities given disease status, as detailed in Table 1. From that, we could construct the optimal ROC curve (Figure 1). The area under the optimal ROC curve is estimated to be 0.580 with Equation 6. The estimated standard error of the AUC is 0.0076 and 0.0075 with Equation 9 and the bootstrap approach, respectively. If we chose 0.0076 as the estimate of the standard error, the corresponding 95% confidence interval (CI) for the AUC is [0.565, 0.595]. The area under the optimal ROC curve has exactly the same value as the one obtained from the logistic regression performed in the original paper.[3] This is not surprising, because only three loci are involved in the test and there is no evidence of interaction among these loci,[3] so that the ROC curve from linear logistic regression should well approximate the optimal ROC curve. Also, from Figure 4 in the original paper, the logistic-regression-based ROC curve is a concave curve that corresponds closely to the optimal ROC curve.

Table 1. Calculation of an Optimal ROC Curve: Three SNPs

Allele Freq.[a] (Case;Control)			Genotype Freq.[b] (Case;Population)			Multilocus Genotype Freq.[c] (Case;Control)		
rs5219	rs1801282	rs7903146	rs5219	rs1801282	rs7903146	rs5219x rs1801282x rs7903146	Likelihood Ratio[d]	Rank Order[e]
0.384;0.354	0.099;0.123	0.384;0.3	0.147;0.126	0.010;0.015	0.147;0.093	0.0002;0.0002	1.236	7
					0.473;0.423	0.0007;0.0006	1.086	11
					0.379;0.484	0.0005;0.0006	0.954	15
				0.178;0.214	0.147;0.093	0.0039;0.0024	1.590	3
					0.473;0.423	0.0124;0.0089	1.394	6
					0.379;0.484	0.0100;0.0082	1.223	9
				0.812;0.771	0.147;0.093	0.0177;0.0086	2.052	1
					0.473;0.423	0.0566;0.0316	1.794	2
					0.379;0.484	0.0454;0.0289	1.570	4
			0.473;0.458	0.010;0.015	0.147;0.093	0.0007;0.0008	0.855	16
					0.473;0.423	0.0022;0.0029	0.753	19
					0.379;0.484	0.0018;0.0027	0.663	23
				0.178;0.214	0.147;0.093	0.0124;0.0114	1.094	10
					0.473;0.423	0.0399;0.0415	0.962	13
					0.379;0.484	0.0320;0.0378	0.846	17
				0.812;0.771	0.147;0.093	0.0566;0.0404	1.402	5
					0.473;0.423	0.1817;0.1476	1.231	8
					0.379;0.484	0.1457;0.1349	1.080	12
			0.379;0.415	0.010;0.015	0.147;0.093	0.0005;0.0009	0.590	24
					0.473;0.423	0.0018;0.0034	0.521	26
					0.379;0.484	0.0014;0.0031	0.459	27
				0.178;0.214	0.147;0.093	0.0100;0.0133	0.753	20
					0.473;0.423	0.0320;0.0483	0.663	22
					0.379;0.484	0.0257;0.0440	0.584	25
				0.812;0.771	0.147;0.093	0.0454;0.0473	0.960	14
					0.473;0.423	0.1457;0.1725	0.845	18
					0.379;0.484	0.1169;0.1572	0.743	21

Details of using the optimal ROC-curve approach for the three SNPs used by Weedon et al.[3]

[a] Allele-frequency estimates obtained from Weedon et al.[3]

[b] Genotype frequencies are computed assuming Hardy-Weinberg equilibrium separately for the case and the control populations.

[c] Multilocus genotype probabilities given disease status calculated with Equation 3.

[d] Likelihood ratios calculated with Equation 4.

[e] Rank order of the LRs.

We can improve the existing genetic test for type 2 diabetes in at least two ways. Clinical studies have shown that diet, physical activity, cigarette smoking, and alcohol consumption affect the risk of type 2 diabetes.[29] These environmental factors could potentially increase the accuracy of the test and, perhaps more importantly, there is the possibility that they interact with the genetic variants to cause the disease.[5,9] With regard to genetic variants, a two-stage genome-wide association study has now been completed for type 2 diabetes.[9] This study confirmed the association with rs7903146 in the TCF7L2 gene and, in addition, seven SNPs were discovered representing four novel disease-susceptibility loci. We combined the information from these new loci, four important environmental factors, and the three variants used in the previous genetic test to create a new predictive genetic test. To avoid overestimating the test's discriminative ability, we chose four SNPs from the seven new SNPs to represent the four novel loci, removing the remaining three SNPs, which are in linkage disequilibrium with the selected loci. The information for these four novel loci, as well as for rs7903146, comes from the confirmatory stage (stage 2)[9] of the study because estimates from that stage are more reliable (i.e., from a well-designed, large-scale association study). Partial details of the calculation are given in Table 2, and we find that the estimated AUC for the new test is 0.657 (Figure 1). In principle, we could also incorporate the three removed SNPs into the tests if the haplotype risk estimates were available, and then the estimated AUC would be even higher.

Since this paper was first written, another genome-wide association study has also been completed, and an additional five novel disease risk loci have been discovered[10] for type 2 diabetes. We therefore obtained the estimates for these five disease risk loci from the second stage of this genome-wide association study and incorporated them also into the new predictive genetic test; partial details of the calculation are given in Table 3. The estimated AUC for the new test is now increased to 0.671 (Figure 1) with an estimated standard error of 0.0071, and the corresponding 95% CI is [0.657, 0.685]. In the near future, more and more disease risk variants will no doubt be discovered,[10] and the relation between these variants (e.g., gene-gene interaction) will become clearer. Thus, for type 2 diabetes, or any other disease, our approach could be adopted to progressively incorporate newly discovered variants, and eventually their interaction effects, gradually improving the classification accuracy of a predictive test.

Compared to the existing genetic test (AUC = 0.580), the proposed test has substantially higher estimated classification

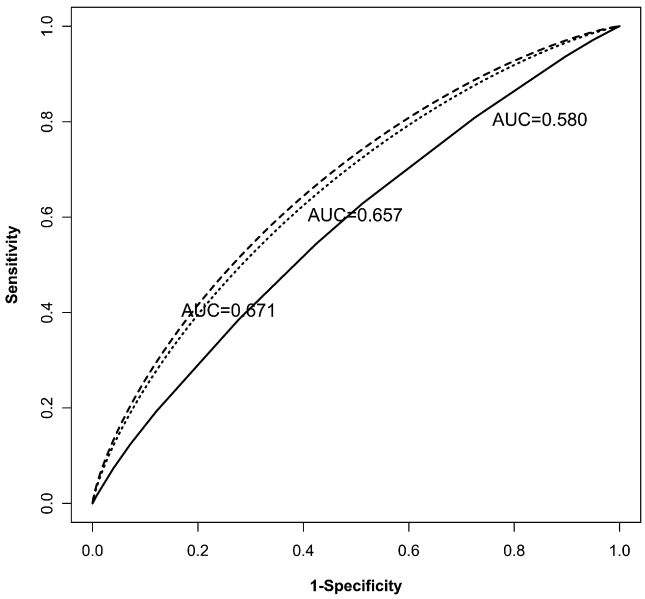

Figure 1. ROC Curves for Type 2 Diabetes
The three lines in the plot from bottom to top correspond to the ROC curves of three type 2 diabetes predictive tests: the rebuilt existing predictive genetic test based on three SNPs, the new predictive test combing the previously associated SNPs, four environmental factors, and four novel risk SNPs from the confirmatory stage of the genome-wide association study,[9] and the improved new predictive test with five additional novel risk SNPs discovered in the second genome-wide association study of type 2 diabetes.[10] The estimated AUC values of these three tests are 0.580, 0.657, and 0.671, respectively.

accuracy (AUC = 0.671). To test the difference in these two AUC values, we calculated

$$\frac{A_1 - A_2}{\sqrt{\mathrm{var}_B(A_1 - A_2)}} = \frac{0.671 - 0.580}{0.008} = 11.375,$$

where A_1 and A_2 are the AUC estimates for the proposed test and existing test, respectively, and $\mathrm{var}_B(A_1 - A_2)$ is the bootstrap variance of $A_1 - A_2$, taking the covariance into account. The corresponding p value for the test is 2.8×10^{-30}. Although this result is exceedingly significant (owing to being based on an asymptotic result and the large samples involved), the AUC estimates and this test made certain assumptions (e.g., a multiplicative model). Therefore, we should design a study to test whether this accuracy can in fact be achieved. The requisite sample size can be calculated to ensure that the new test has a higher level of accuracy than the existing test. By using Equation 10, we can calculate the variances: $\mathrm{var}_D = 0.072$ and $\mathrm{var}_{\overline{D}} = 0.074$. Assuming a type I error rate (α) of 0.05, power ($1 - \beta$) of 0.95, and equal numbers of cases and controls ($p = 1$), we compute the necessary sample size using the general formula (Equation 12),

$$n_D = n_{\overline{D}} = (0.072 + 0.074)\left\{\frac{1.645 + 1.645}{0.671 - 0.580}\right\}^2 \approx 191,$$

i.e., 191 cases and 191 controls. Note that this particular sample size also applies for testing our hypothesis with any choices of type I error rate and power that satisfy $\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta) = 1.645 + 1.645 = 3.29$ (e.g., $1 - \beta = 0.8$ and $\alpha = 0.007$).

Discussion

The importance and benefit of predictive genetic tests have been recognized by both researchers and the public.[7,8,30–33] In a recent article, genetic tests have been described as the cornerstone of genomic medicine.[7] Although the benefit of predictive genetic testing is obvious and this opportunity is important for the field of medical genetics, work on this topic is still limited. This is partly due to our limited knowledge of genetic causes of common diseases. With the recent intensive research on common diseases, especially with the completion of genome-wide association studies, many novel, apparently causal genes have already been discovered, and variants of these genes, whether themselves causal or not, could be usefully implemented into genetic tests. For that purpose, we have described here a general model to both design and evaluate a predictive genetic test. By taking the information from previous, related association studies, our approach has the ability to estimate the proposed test's approximate discriminative ability. By using this result, if it is encouraging, we can formulate a hypothesis of interest to rigorously evaluate the proposed test, and for this purpose, we have provided the formula for sample-size calculation. Our approach is easy to use. In a simple scenario (e.g., for the rebuilt existing predictive genetic test for type 2 diabetes), it can be implemented with Excel. We have also provided sample R source code (OPMDesign.R) on the website noted in the Web Resources.

The model we have introduced for the design of a predictive genetic test is an illustration of the use of the original optimality theory based on the likelihood ratio.[19,20] This theory indicates that a decision rule based on the likelihood ratio is best. We have further shown that a test built on the likelihood ratio can achieve the highest discriminative ability among all approaches. We incorporated these ideal properties in designing a predictive genetic test for type 2 diabetes. Our approach is similar to the approaches introduced by Baker[22] and McIntosh et al.[21] The differences among the three approaches relate to the calculation of the LR, or its one-to-one function, the risk score. Baker[22] approached the LR by directly using the joint distributions of predictors among cases and controls in the data, whereas McIntosh et al.[21] estimated the risk by logistic regression. From a large-sample simulation[21] and a real data application,[22] Baker's approach showed a better performance than did the logistic-regression approach,[21] but with an over-fitting issue when the sample size is limited. A model is liable to over-fitting whenever too many parameters are estimated for the sample size that is available.[34] Baker approached the LR directly from the joint distributions of the predictors, with the result that the model complexity increases exponentially with the number of predictors, leading to an overly optimistic estimate of the ROC performance. For example, if the test is based on n risk SNPs, this approach requires $3^n - 1$ genotype combination frequencies to be estimated from each of the case and

Table 2. Calculation of an Optimal ROC Curve: Adding Four SNPs—One Shown—and Four Environmental Factors—One Shown—to the Calculations in Table 1

Information for Environmental Risk Factors[a] (Percentage; Relative Risk)		Allele Freq.[b] (Case; Control)		Genotype Freq.[c] (Case; Control)		Frequencies of Environmental[d] and Genetic Risk Factors (Case; Population)						Environmental and Genetic Risk Factors Combination Frequencies (Case;Control)		
Dietary Score	…	rs5219	…	rs7903146	…	Dietary Score	…	rs5219	…	rs7903146	…	Dietary Score x … x rs5219x …	Likelihood Ratio	Rank Percentage[e]
0.150;1.000	…	0.384;0.354	…	0.163; 0.084	…	0.197;0.150	…	0.147;0.126	…	0.163;0.089	…	8.238e-09;1.450e-09	5.680	0.20%
	…	:	…	…	…	:	…	…	…	…	…	…	…	…
0.270;0.860	…	0.384;0.354	…	0.163; 0.084	…	0.305;0.270	…	0.147;0.126	…	0.163;0.089	…	1.275e-08;2.767e-09	4.608	0.50%
	…	:	…	…	…	:	…	…	…	…	…	…	…	…
0.170;0.770	…	0.384;0.354	…	0.163; 0.084	…	0.172;0.170	…	0.147;0.126	…	0.163;0.089	…	7.189e-09;1.806e-09	3.982	0.90%
	…	:	…	…	…	:	…	…	…	…	…	…	…	…
0.260;0.670	…	0.384;0.354	…	0.163; 0.084	…	0.229;0.260	…	0.147;0.126	…	0.163;0.089	…	9.568e-09;2.869e-09	3.335	1.75%
	…	:	…	…	…	:	…	…	…	…	…	…	…	…
0.150;0.490	…	0.384;0.354	…	0.163; 0.084	…	0.097;0.150	…	0.147;0.126	…	0.163;0.089	…	4.037e-09;1.767e-09	2.285	5..86%
	…	:	…	…	…	:	…	…	…	…	…	…	…	…

Details of using the optimal ROC-curve approach to combine the SNPs in Table 1, four environmental factors, and four novel risk SNPs from the first genome-wide association study[9] by assuming $\rho = 0.07$ (five combinations only, out of a total of 9×10^5 combinations).

[a] Distribution and relative risk of the environmental risk factors obtained from Hu et al.[29]

[b] Allele frequency estimates obtained from Weedon et al.[3]

[c] Genotype frequency estimates for cases and controls obtained from the confirmatory stage (stage 2) of the first genome-wide association study.[9]

[d] Distribution of the environmental factors in cases calculated by applying Equation 1.

[e] 100•(rank of LR)/(total number of combinations).

control samples. Our approach approximates the likelihood ratios by utilizing the essential information of each genetic variant from previous genetic studies and assumes a multiplicative model. This not only allows us full use of the information for single genetic variants, which is commonly obtained in genetic studies, but also helps mitigate against over-fitting. For the same example, our approach needs only $2n$ genotype frequencies to be estimated from each of the case and control samples to compute the frequencies of all genotype combinations, and therefore the model complexity is greatly reduced. Our approach can use genetic variants with known mode of inheritance and is easily extended to scenarios where the genetic variants are in linkage disequilibrium or interact with each other. Incorporating this additional information can make the approach more robust to the over-fitting problem and increase the power of the approach. However, the performance of the approach relies on the assumptions

Table 3. Calculation of an Optimal ROC Curve: Adding Five SNPs—One Shown—to the Calculations in Table 2

Environmental and Genetic Risk Factors Combination Frequencies from Previous Test[a] (Case;Control)		Genotype Freq.[b] (Case;Control)		Genotype Freq. (Case;Population)		Environmental and Genetic Risk Factors Combination Frequencies (Case;Control)	Likelihood Ratio	Rank Percentage[c]
Dietary Score x … x rs5219x …		rs4402960	…	rs4402960	…	Dietary Score x … x rs4402960 x …		
8.238e-09;1.450e-09	…	0.103;0.115	…	0.103;0.114	…	1.118e-11;1.160e-12	9.640	0.02%
…	…	…	…	…	…	…	…	…
1.275e-08;2.767e-09	…	0.103;0.115	…	0.103;0.114	…	1.731e-11;2.300e-12	7.523	0.05%
…	…	…	…	…	…	…	…	…
7.189e-09;1.806e-09	…	0.103;0.115	…	0.103;0.114	…	9.756e-12;1.534e-12	6.359	0.10%
…	…	…	…	…	…	…	…	…
9.568e-09;2.869e-09	…	0.103;0.115	…	0.103;0.114	…	1.298e-11;2.492e-12	5.210	0.21%
…	…	…	…	…	…	…	…	…
4.037e-09;1.767e-09	…	0.103;0.115	…	0.103;0.114	…	5.478e-12;1.589e-12	3.447	0.89%
…	…	…	…	…	…	…	…	…

Details of using the optimal ROC-curve approach to further improve the new predictive genetic test in Table 2 by incorporating five additional novel risk SNPs discovered in the second genome-wide association study of type 2 diabetes[10] (five combinations only, out of a total of 2×10^8 combinations).

[a] Multilocus genotype probabilities given disease status obtained from the new predictive genetic test in Table 2.

[b] Genotype frequency estimates obtained from the confirmatory stage (stage 2) of the second genome-wide association study.[10]

[c] 100•(rank of LR)/(total number of combinations).

being satisfied and the accuracy of the information from previous genetic studies. Assumption violation and inaccurate information can bias the parameters of interest (e.g., the AUC). For instance, if we violate the assumption of a multiplicative model and linkage equilibrium by incorporating all seven SNPs discovered from the first genome-wide association study into the test, the estimated AUC will increase from 0.671 to 0.678. Although this does not seem to affect the result too much, violation of the assumptions when there is a large number of loci or loci having strong effects (i.e., high relative risks) will cause serious bias. In illustration of this, assume two loci near each risk locus are also used for forming the test. If we assume for simplicity that the two loci are in complete linkage disequilibrium with the risk loci, then the estimated AUC is increased to 0.750. To avoid introducing such a bias, we can either apply the extended approach to incorporate these loci given the required information (i.e., the LD estimates), if available, or we can simply remove these extra loci—this leads to a conservative estimate, and the design will still be valid.

We illustrated the proposed approach for the case of type 2 diabetes. The approach was first examined with as an example an existing predictive genetic test,[3] and the result from the optimal ROC curve method was found to be highly consistent with the one originally reported. This results from the equivalence between the logistic-regression-based ROC curve and the optimal ROC curve in such a simple scenario (i.e., few loci, no interactions). To further improve the predictive test, we incorporated further risk factors. With both important environmental factors and novel loci discovered from two recent genome-wide association studies taken into account, the new predictive genetic test (AUC = 0.671) could have a significantly higher classification accuracy ($P = 2.8 \times 10^{-30}$) than the existing test (AUC = 0.580). Because the variants involved in the new test are either well studied or confirmed, and the corresponding estimates used came from well-designed, large-scale studies, the estimated AUC value for this new test could be considered to be a reasonable approximation of the actual classification accuracy of the test. Because gene-gene and gene-environmental interactions were not studied in the previous association studies, we are unable to incorporate these effects into the new proposed test. If there are strong interaction effects among the predictors, our estimated AUC value would tend to be conservative. The design to study the test will still be valid because any strong interaction effect would lead to a higher value of the AUC and thus be in favor of the alternative hypothesis.

Unlike the current predictive genetic test, we also incorporated environmental risk factors into the new test—not only because by themselves they can increase information on risk to the disease, but also because they could interact with the genetic variants to cause the disease. Without considering them, we cannot study any gene-environmental effect in any proposed new test. The other advantage of studying them is that in some scenarios we can use them as a method to help disease prevention. For example, we could use the equation in Appendix C to calculate the positive predictive values (PPV) for individuals who carry the multilocus genotype with most risk. The chance of type 2 diabetes would then be predicted to decrease from 83.9% to 29.2% if they adopted a healthy life style (exercise/week $>$ 7 hr and dietary score = 5) rather than having a nonhealthy life style (exercise/week $<$ 0.5 hr and dietary score = 1). Because the predictive genetic test can be conducted at an early age, such as at birth, it would be relatively easy to advise high-risk people to adopt a healthier life style when they are young rather than make them change behavior after the disease has been diagnosed.

Some researchers have suggested that it would be less costly and more efficient to conduct genetic tests for only high-risk individuals (e.g., individuals with a family history of disease), instead of for the general population.[35,36] We therefore also attempted to investigate a predictive test for high-risk diabetes subjects on the basis of results of the initial stage of the first genome-wide association study.[9] We found that the resulting test could reach a high level of classification accuracy (AUC = 0.855). However, this result would be liberal because the controls came from the general population rather than from a subpopulation of high-risk individuals. To conduct a genetic test on a particular subgroup of the population requires that we investigate its performance on that subgroup. If a test appears to be superior to the one already in use for the general population, it might be considered as a candidate for a high-risk population, but it must first be carefully tested on such a population.

Our proposed approach should be a useful tool for designing a predictive genetic test. It would help the investigator explore possible hypotheses and make decisions regarding developing a new genetic test. It should function as an exploratory phase of medical test development at little cost. The performance of the approach depends on our knowledge of disease-associated variants and the accuracy of the estimates found in the published association studies. The estimated test's classification accuracy will reflect the actual test's performance if the variants involved in the test have been well studied and their estimates come from well-planned studies. Otherwise, the estimate could be subject to bias. Any result would only be valid for the same population as that used for the association study that produced the estimates, and not necessarily apply to different populations, for which the risk estimates and population frequency estimates could be different.

Appendix A

Proof that the Optimal ROC Curve Has the Highest AUC

When we are dealing with multiple genetic variants, each method combines multiple predictors differently and assigns its own unique score for each multilocus genotype.

Because the ROC curve relies only on the ranks of these scores, not the absolute scores, the ROC curve from each approach can be represented by the unique ranks of the multilocus genotypes. Assuming o_1 represents the ranks of multilocus genotypes from the optimal ROC approach and o_2 represents those from any other approach, we prove that the AUC from the optimal ROC approach (AUC^{O_1}) is always as great as, or greater than, that from any other approach (AUC^{O_2}).

It is easy to show that the rank o_2 can always be obtained from o_1 by a series of order switches between pairs of multilocus genotypes. We prove that each order switch from the original ranks of the optimal ROC curve can only decrease, or leave unchanged, the AUC value. Assume $P^n = (p^n_{(1)}, p^n_{(2)}, \cdots, p^n_{(K)})$ and $Q^n = (q^n_{(1)}, q^n_{(2)}, \cdots, q^n_{(K)})$ are the distributions of the multilocus genotypes in cases and controls, respectively, from the nth order switch. At the $(n+1)$th step, we switch the order of the ith and jth multilocus genotypes, and the corresponding distributions are then denoted $P^{n+1} = (p^n_{(1)}, \cdots, p^n_{(j)}, \cdots, p^n_{(i)}, \cdots, p^n_{(K)})$ and $Q^{n+1} = (q^n_{(1)}, \cdots, q^n_{(j)}, \cdots, q^n_{(i)}, \cdots, q^n_{(K)})$. Simply by using the trapezoid rule, we can calculate the difference in the AUCs:

$$
\begin{aligned}
&AUC^n - AUC^{n+1}\\
&= \frac{1}{2}(2p^n + p^n_{(i)})q^n_{(i)} + \frac{1}{2}(2p^n + 2p^n_{(i)} + p^n_{(i+1)})q^n_{(i+1)}\\
&\quad + \cdots + \frac{1}{2}\left(2p^n + 2\sum_{k=i}^{j-2} p^n_{(k)} + p^n_{(j-1)}\right)q^n_{(j-1)}\\
&\quad + \frac{1}{2}\left(2p^n + 2\sum_{k=i}^{j-1} p^n_{(k)} + p^n_{(j)}\right)q^n_{(j)}\\
&\quad - \frac{1}{2}(2p^n + p^n_{(j)})q^n_{(j)} - \frac{1}{2}(2p^n + 2p^n_{(j)} + p^n_{(i+1)})q^n_{(i+1)}\\
&\quad - \cdots - \frac{1}{2}\left(2p^n + 2p^n_{(j)} + 2\sum_{k=i+1}^{j-2} p^n_{(k)} + p^n_{(j-1)}\right)q^n_{(j-1)}\\
&\quad - \frac{1}{2}\left(2p^n + 2p^n_{(j)} + 2\sum_{k=i+1}^{j-1} p^n_{(k)} + p^n_{(i)}\right)q^n_{(i)}\\
&= \frac{1}{2}(p^n_{(i)}q^n_{(i)} - p^n_{(j)}q^n_{(j)}) + (p^n_{(i)} - p^n_{(j)})q^n_{(i+1)} + \cdots\\
&\quad + (p^n_{(i)} - p^n_{(j)})q^n_{(j-1)} + \frac{1}{2}(p^n_{(j)}q^n_{(j)} - p^n_{(i)}q^n_{(i)})\\
&\quad + \left(\sum_{k=i}^{j-1} p^n_{(k)}\right)q^n_{(j)} - \left(p^n_{(j)} + \sum_{k=i+1}^{j-1} p^n_{(k)}\right)q^n_{(i)}\\
&= (p^n_{(i)}q^n_{(j)} - p^n_{(j)}q^n_{(i)}) + (p^n_{(i)}q^n_{(i+1)} - p^n_{(i+1)}q^n_{(i)})\\
&\quad + (p^n_{(i+1)}q^n_{(j)} - p^n_{(j)}q^n_{(i+1)}) + \cdots + (p^n_{(i)}q^n_{(j-1)} - p^n_{(j-1)}q^n_{(i)})\\
&\quad + (p^n_{(j-1)}q^n_{(j)} - p^n_{(j)}q^n_{(j-1)}),
\end{aligned}
$$

where

$$p^n = \sum_{k=1}^{i-1} p^n_{(k)}.$$

Because $\frac{p^n_{(i)}}{q^n_{(i)}} \geq \frac{p^n_{(i+1)}}{q^n_{(i+1)}} \geq \cdots \geq \frac{p^n_{(j)}}{q^n_{(j)}}$,

$$AUC^n \geq AUC^{n+1}\left(AUC^n = AUC^{n+1} \text{ if and only if } \frac{p^n_{(i)}}{q^n_{(i)}} = \frac{p^n_{(i+1)}}{q^n_{(i+1)}} = \cdots = \frac{p^n_{(j)}}{q^n_{(j)}}\right).$$

To satisfy the condition $LR^n_{(i)} \geq LR^n_{(i+1)} \geq \cdots \geq LR^n_{(j)}$ ($LR^n_{(k)} = p^n_{(k)}/q^n_{(k)}, k = i, \cdots, j$), the order of the ith to jth multilocus genotypes must keep their original order as in o_1. This requirement also applies to other pair switches and may limit the possible order changing but is always feasible. Therefore, we obtain

$$AUC^{O_1} \geq \cdots \geq AUC^n \geq AUC^{n+1} \geq \cdots \geq AUC^{O_2},$$

and thus prove that the AUC of the optimal ROC curve is at worst equal to that of any other approach. Because O_2 is arbitrary, the AUC of the optimal ROC curve is the highest among all ROC curves.

Appendix B

Risk Estimates Measured as Odds Ratios

If the estimates of risk parameters for the genetic variants are odds-ratio estimates, Equation 1 is no longer valid unless we make the rare disease assumption. Under the common disease scenario, we could still obtain $P(g_{ij_i}|D)$, $j_i = 1, \cdots, m_i$, assuming there are m_i possible (multilocus) genotypes for the ith genetic variant. For each genotype, we denote its odds ratio estimate $OR_{ij_i} = P(D|g_{ij_i})/P(\overline{D}|g_{ij_i})/[P(D|g_{i1})/P(\overline{D}|g_{i1})]$ and the corresponding frequency f_{ij_i}, $j_i = 1, \cdots, m_i$. The probability of disease given genotype, $P(D|g_{ij_i})$, $j_i = 1, \cdots, m_i$, can then be obtained from the following m_i equations,

$$\begin{cases} P(D\,|\,g_{ij_i}) = \dfrac{OR_{ij_i}P(D\,|\,g_{i1})}{1 + OR_{ij_i}P(D\,|\,g_{i1}) - P(D\,|\,g_{i1})}, j_i = 1, \cdots, m_i. \\ \sum_{j_i=1}^{m_i} P(D\,|\,g_{ij_i})f_{ij_i} = \rho \end{cases} \quad (13)$$

By applying Bayes' rule, we have $P(g_{ij_i}|D)$,

$$P(g_{ij_i}|\,D) = \frac{P(D\,|\,g_{ij_i}) \cdot f_{ij_i}}{\rho},\ j_i = 1, \cdots, m_i. \quad (14)$$

Genetic Variants Are SNPs

We assume the ith genetic variant is a SNP that has two alleles, A and a. From a previous study, we obtain the genotype frequencies $F = (f_2, f_1, f_0)$ for genotypes $G = (AA, Aa, aa)$ and the relative risk rr_2 (rr_1), the risk for AA (Aa) divided by that for aa. If the genotype frequencies are not available from the previous study, we can estimate them from the allele frequencies on the assumption of HWE. On the basis

of Equation 1, we have the genotype distribution in the disease population,

$$\begin{cases} P(AA\,|\,D) = \dfrac{rr_2 f_2}{rr_2 f_2 + rr_1 f_1 + f_0} \\ P(Aa\,|\,D) = \dfrac{rr_1 f_1}{rr_2 f_2 + rr_1 f_1 + f_0} \\ P(aa\,|\,D) = \dfrac{f_0}{rr_2 f_2 + rr_1 f_1 + f_0} \end{cases}. \quad (15)$$

Further modifying the above equations, we can have a similar expression for a different mode of inheritance. For instance, if the variant A is dominant, we have the following conditional probabilities:

$$\begin{cases} P(AA,Aa\,|\,D) = \dfrac{rr(f_2 + f_1)}{rr(f_2 + f_1) + f_0} \\ P(aa\,|\,D) = \dfrac{f_0}{rr(f_2 + f_1) + f_0} \end{cases}, \quad (16)$$

where *rr* denote the relative risk of the genotypes with the A allele versus the genotype without the A allele.

Using Information from Previous Genetic-Test Studies

If the genetic variants we are interested in have been studied in an existing genetic test, we could also utilize such information. From the ROC curve of the previous genetic test, we can obtain the entire set of TPR and FPR pairs. Assuming there are K TPR and FPR pairs, the distribution of genotype combinations among cases can be derived as

$$P(G_k\,|\,D) = TPR_{(k+1)} - TPR_{(k)}, k = 1,\cdots,K-1\,,$$

where the $TPR_{(k)}$ are the ordered TPRs from the left side to the right of the ROC curve.

If we are only interested in some of the genetic variants, we sum the above conditional probabilities over all genetic variants of no interest and thus obtain the distribution for the genetic variants of interest.

Appendix C

Calculating Predictive Values

For given disease prevalence ρ, the predictive values can be simply calculated by Bayes' rule:

$$\begin{cases} PPV_{(k)} = \dfrac{TPR_{(k)} \cdot \rho}{TPR_{(k)} \cdot \rho + FPR_{(k)} \cdot (1-\rho)} \\ NPV_{(k)} = \dfrac{(1 - FPR_{(k)}) \cdot (1-\rho)}{(1 - FPR_{(k)}) \cdot (1-\rho) + (1 - TPR_{(k)}) \cdot \rho} \end{cases}, k = 1,\cdots,K.$$

Acknowledgments

This work was supported by United States Public Health Service Resource grant (RR03655) from the National Center for Research Resources, Research grant (GM28356) from the National Institute of General Medical Sciences, and Cancer Center Support Grant P30CAD43703 from the National Cancer Institute.

Received: September 22, 2007
Revised: November 16, 2007
Accepted: December 20, 2007
Published online: March 6, 2008

Web Resources

The URLs for data presented herein are as follows:

R code, http://darwin.cwru.edu/~qlu/

References

1. Christensen, K., and Murray, J.C. (2007). What genome-wide association studies can do for medicine. N. Engl. J. Med. *356*, 1094–1097.
2. Yang, Q., Khoury, M.J., Botto, L., Friedman, J.M., and Flanders, W.D. (2003). Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. Am. J. Hum. Genet. *72*, 636–649.
3. Weedon, M.N., McCarthy, M.I., Hitman, G., Walker, M., Groves, C.J., Zeggini, E., Rayner, N.W., Shields, B., Owen, K.R., Hattersley, A.T., et al. (2006). Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. PLoS Med. *3*, e374.
4. Yiannakouris, N., Trichopoulou, A., Benetou, V., Psaltopoulou, T., Ordovas, J.M., and Trichopoulos, D. (2006). A direct assessment of genetic contribution to the incidence of coronary infarct in the general population Greek EPIC cohort. Eur. J. Epidemiol. *21*, 859–867.
5. Lyssenko, V., Almgren, P., Anevski, D., Perfekt, R., Lahti, K., Nissen, M., Isomaa, B., Forsen, B., Homstrom, N., Saloranta, C., et al. (2005). Predictors of and longitudinal changes in insulin sensitivity and secretion preceding onset of type 2 diabetes. Diabetes *54*, 166–174.
6. Janssens, A.C., and van Duijn, C.M. (2006). Towards predictive genetic testing of complex diseases. Eur. J. Epidemiol. *21*, 869–870.
7. Epstein, C.J. (2006). Medical genetics in the genomic medicine of the 21st century. Am. J. Hum. Genet. *79*, 434–438.
8. Evans, J.P., Skrzynia, C., and Burke, W. (2001). The complexities of predictive genetic testing. BMJ *322*, 1052–1056.
9. Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature *445*, 881–885.
10. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science *316*, 1341–1345.
11. Bookman, E.B., Langehorne, A.A., Eckfeldt, J.H., Glass, K.C., Jarvik, G.P., Klag, M., Koski, G., Motulsky, A., Wilfond, B., Manolio, T.A., et al. (2006). Reporting genetic results in research studies: summary and recommendations of an NHLBI working group. Am. J. Med. Genet. A. *140*, 1033–1040.

12. Zweig, M.H., and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin. Chem. *39*, 561–577.
13. Lusted, L.B. (1971). Signal detectability and medical decision-making. Science *171*, 1217–1219.
14. Janssens, A.C., Pardo, M.C., Steyerberg, E.W., and van Duijn, C.M. (2004). Revisiting the clinical validity of multiplex genetic testing in complex diseases. Am. J. Hum. Genet. *74*, 585–588.
15. Yang, Q., Khoury, M.J., Botto, L., Friedman, J.M., and Flanders, W.D. (2004). Revisiting the clinical validity of multiplex genetic testing in complex diseases: Reply to Janssens et al. Am. J. Hum. Genet. *74*, 588–589.
16. McClish, D.K. (1989). Analyzing a portion of the ROC curve. Med. Decis. Making *9*, 190–195.
17. Hanley, J.A. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. Crit. Rev. Diagn. Imaging *29*, 307–335.
18. Pepe, M.S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. Biometrics *62*, 221–229.
19. Green, D.M., and Swets, J.A. (1966). Signal Detection Theory and Psychophysics (New York: John Wiley and Sons).
20. Egan, J.P. (1975). Signal Detection Theory and ROC Analysis (New York: Academic Press).
21. McIntosh, M.W., and Pepe, M.S. (2002). Combining several screening tests: Optimality of the risk score. Biometrics *58*, 657–664.
22. Baker, S.G. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. Biometrics *56*, 1082–1087.
23. Weinstein, M.C., Fineberg, H.V., Elstein, A.S., Frazier, N.S., Neuhauser, D., Neutra, R.R., and McNeil, B.J. (1980). Clinical Decision Analysis (Philadelphia: W.B. Saunders).
24. Gorelick, R., and Laubichler, M.D. (2004). Decomposing multilocus linkage disequilibrium. Genetics *166*, 1581–1583.
25. Marchini, J., Donnelly, P., and Cardon, L.R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat. Genet. *37*, 413–417.
26. DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics *44*, 837–845.
27. Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., Bailey-Wilson, J.E., et al. (2007). Replicating genotype-phenotype associations. Nature *447*, 655–660.
28. Pepe, M.S. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction (New York: Oxford University Press).
29. Hu, F.B., Manson, J.E., Stampfer, M.J., Colditz, G., Liu, S., Solomon, C.G., and Willett, W.C. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. N. Engl. J. Med. *345*, 790–797.
30. Beaudet, A.L. (1999). 1998 ASHG presidential address. Making genomic medicine a reality. Am. J. Hum. Genet. *64*, 1–13.
31. Bell, J. (1998). The new genetics in clinical practice. BMJ *316*, 618–620.
32. Collins, F.S. (1999). Shattuck lecture–medical and societal consequences of the Human Genome Project. N. Engl. J. Med. *341*, 28–37.
33. Jones, M. (2000). The genetic report card. *New York Times Magazine*, June 11, p. 80.
34. Hastie, T., Tibshirani, R., and Friedman, J.H. (2001). The Elements of Statistical Learning (New York: Springer).
35. Khoury, M.J. (2003). Genetics and genomics in practice: The continuum from genetic disease to genetic information in health and disease. Genet. Med. *5*, 261–268.
36. Vineis, P., Schulte, P., and McMichael, A.J. (2001). Misconceptions about the use of genetic tests in populations. Lancet *357*, 709–712.