Statistics and Its Interface Volume 4 (2011) 373–380

A clustered optimal ROC curve method for family-based genetic risk prediction

Chengyin Ye, Jun Zhu* and Qing Lu*

Risk prediction that capitalizes on emerging genetic findings holds great promises for improving public health and clinical care. Statistical methods for genetic risk prediction research, and particularly for correlated data, are however still lacking. To address this, we have developed a clustered optimal ROC curve (CORC) method, in order to build predictive genetic tests using data from family-based genetic research. For the proposed method, we have extended the conventional optimal ROC curve method to handle multiple genetic markers, taking sample correlation into consideration, and implemented a forward selection algorithm to allow for high-dimensional data and the capture of possible epistasis. We have evaluated the CORC method using both simulations and a real-data application, showing that the method performed better than other existing methods under various pedigree structures and underlying disease models. In the real-data application, we applied the method to the large scale International Multi-Center ADHD Genetics Project dataset and formed a predictive genetic test for conduct disorder. The test reached a low to medium classification accuracy, with an AUC value of 0.6908.

Keywords and phrases: Clustered ROC curve, Predictive genetic test, High-dimensional data, Genome-wide association study.

1. INTRODUCTION

The completion of the Human Genome and International HapMap Projects brought us a new tool for increasing our understanding of common complex diseases, as well as an opportunity for early disease prediction and prevention [1]. By combining multiple genetic risk variants from recent genetic research, as well as existing clinical risk factors, predictive genetic testing has considerable potential for accurate risk assessment and for use in screenings and prevention [2]. If it succeeds, predictive genetic testing would play an important role in shifting the focus of medical care from intervention to prevention [3], and could eventually lead to reduced morbidity and mortality [2].

In previous research, predictive genetic tests have been constructed not only with monogenetic disorders, such as Huntington's disease [4], but also for complex diseases, such as cardiovascular disease [5,6] and type 2 diabetes (T2D) [7,8]. These have revealed that integrating dozens of common risk variants from genome-wide association studies (GWAS) could form more accurate predictive genetic tests for complex diseases [9,10]. In risk prediction research, a genotype-score-based method has commonly been used to assess an individual's disease risk [5,8,10,11]. This genotype-scoring method derives a global genotype risk score by counting the number of risk alleles over multiple genes, and then uses this global genotype risk score to predict an individual's disease risk. The genotype-scoring method is convenient and easy to interpret. However, it usually requires assumptions of equal effect sizes. In order to address this limitation, weighting approaches have been proposed that would incorporate the estimated effect sizes as weights into the genotype-scoring method [12]. The weighting approaches relax the assumption of equal effect sizes. We would caution, however, that the empirical weighting processes could introduce bias if there are variations across studies or insufficient information in previous studies for reliable parameter estimation [13]. Moreover, the genotype-scoring method assumes that all of the testing variants are disease-related and are independent of each other. It is, therefore, subject to low performance when non-causal genetic variants or interactions are present.

Family-based design is one of the most popular study designs in genetic research today. Data generated from these studies are a valuable resource for genetic risk prediction research. Using the existing methods, family-based risk prediction research has been conducted to assess the combined effect of multiple risk variants in disease prediction. For example, James B. Meigs et al. implemented a family-based genotype scoring method—a pooled logistic regression model with generalized estimating equations (GS-GEE), fitted on the global genotype risk score—to investigate a predictive genetic test for T2D [7]. GS-GEE could be easily applied to a large number of genetic risk variants, taking sample correlations into consideration, which would make it suitable for family-based risk prediction.

There is also a long history of investigating genetic prediction in inbred plant and animal populations. Although the focus here is on human genetics, many of the concepts and approaches used within the context of animal and plant breeding have the potential for applications in human genetics. For example, penalized regression methods, such as

*Corresponding authors.

the Bayesian Lasso method, have been proposed for predicting quantitative traits in animal and plant breeding studies. These methods commonly evaluate the predictive ability of high-dimensional marker sets using pedigree data, and are, therefore, strongly related to methods used in family-based risk prediction [14,15]. The problem with both the Bayesian Lasso and the GS-GEE methods is that they are not suitable for detecting interactions, particularly high order interactions, on high-dimensional marker sets, and, thus, could be subject to low performance when interactions exist.

We propose a nonparametric method—a clustered optimal ROC curve method (CORC)—for family-based genetic risk prediction research. The CORC method requires no assumptions of equal effect size or independence, and thus could have a more robust performance. It combines the strength of the optimal ROC curve [16,17,18] and the clustered ROC curve, and could theoretically form a test with the highest classification accuracy for family data. Moreover, the CORC method implements a computationally efficient algorithm—the forward selection algorithm—for the variable selection process, which makes the method feasible for dealing with a large number of loci, taking possible interactions into consideration. The proposed CORC method is capable of handling arbitrary pedigree structures, and can thus be used for family studies with different pedigree structures, including simple pedigrees with two generations, complex pedigrees with several generations, or a mixture of different pedigrees. Through simulation studies and a real-data application of conduct disorder (CD) disease, we compared the proposed CORC method with both the GS-GEE and the forward ROC curve methods, where predictive genetic tests for CD were evaluated using data from the International Multi-Center ADHD Genetics Project.

2. THE CORC METHOD

The ROC curve is widely used to measure a test's overall accuracy. It plots a test's sensitivity vs. specificity by continually varying the threshold of the test results [19]. When varying the likelihood ratios (LRs) of test results from the largest to the smallest value, the ROC curve can attain its optimality, resulting in an optimal ROC curve [20]. The test formed by the optimal ROC curve has many ideal properties at each point on the curve; e.g., for a fixed value of specificity (sensitivity), it has the highest sensitivity (specificity). We have incorporated the concept of the optimal ROC into our proposed CORC method.

We assume a total of M individuals from N families (i.e., cluster), each with measurements of p predictors (e.g., SNPs) and a binary response (e.g., disease status). Let y_{ij} ($y_{ij} \in S$) denote the binary response of the jth individual ($j = 1, 2, \ldots, m_i$) in the ith family ($i = 1, 2, \ldots, N$), which has two possible values: $S = 1$ (e.g., disease) and $S = 0$ (e.g., non-disease). Using $G_{ij}^p = (g_{ij1}, g_{ij2}, \ldots, g_{ijp})$ ($G_{ij}^p \in G_t^p$), we denote the measurements of p predictors for the jth individual in the ith family, which belongs to one of p-dimensional risk profiles, G_t^p ($t = 1, 2, \ldots, p_s$), where $G_t^p = (g_{1t}, g_{2t}, \ldots, g_{pt})$. The conditional distributions of $G_t^p, P(G_t^p|S)$, can be calculated as

$$(1) \qquad P(G_t^p|S) = \frac{\sum_i^N \sum_j^{m_i} I_{\{(i,j):G_{ij}^p = G_t^p, y_{ij}=S\}}(i,j)}{\sum_i^N \sum_j^{m_i} I_{\{(i,j):y_{ij}=S\}}(i,j)} \qquad t = 1, \ldots, p_s, \quad S = 0, 1,$$

where

$$I_{\{(i,j):G_{ij}^p = G_t^p, y_{ij}=S\}}(i,j) = \begin{cases} 1 \text{ if } G_{ij}^p = G_t^p \text{ and } y_{ij} = S \\ 0 \text{ if } G_{ij}^p \neq G_t^p \text{ or } y_{ij} \neq S \end{cases},$$

$$I_{\{(i,j):y_{ij}=S\}}(i,j) = \begin{cases} 1 \text{ if } y_{ij} = S \\ 0 \text{ if } y_{ij} \neq S \end{cases}.$$

Given the conditional probability, $P(G_t^p|S)$, we obtain the LR of G_t^p by using $LR(G_t^p) = P(G_t^p|S = 1)/P(G_t^p|S = 0)$. If samples in the study are independent, we then rank the individuals' LRs from the largest to the smallest value and plot the optimal ROC curve. The area under the ROC curve (AUC), the most popular one-dimensional index of the ROC curve, could be summarized to represent the overall test's classification accuracy. However, this approach cannot be used directly on family data since individuals from the same family are related. To take the sample correlation into account, we have incorporated a clustered ROC curve method, originally proposed by Obuchowski [21]. For a particular family i, let $LR^a(G_{il}^p)$ denote the likelihood ratio (LR) of the lth affected individual ($l = 1, \ldots, a_i$) carrying a p-dimensional risk profile, G_{il}^p, and let $LR^u(G_{ik}^p)$ denote that of the kth unaffected individual ($k = 1, \ldots, u_i$) carrying a p-dimensional risk profile, G_{ik}^p, where a_i and u_i represent the total number of affected and unaffected individuals in family $i (a_i + u_i = m_i)$, respectively. We then calculate the clustered AUC,

$$(2) \quad AUC_c^p = \frac{1}{AU} \sum_{i=1}^{N} \sum_{i'=1}^{N} \sum_{l=1}^{a_i} \sum_{k=1}^{u_{i'}} \varphi(LR^a(G_{il}^p), LR^u(G_{i'k}^p)),$$

where

$$A = \sum_i^N a_i, \quad U = \sum_i^N u_i,$$

and

$$\varphi(LR^a(G_{il}^p), LR^u(G_{i'k}^p)) = \begin{cases} 1.0 \text{ if } LR^a(G_{il}^p) > LR^u(G_{i'k}^p) \\ 0.5 \text{ if } LR^a(G_{il}^p) = LR^u(G_{i'k}^p) \\ 0.0 \text{ if } LR^a(G_{il}^p) < LR^u(G_{i'k}^p) \end{cases}.$$

Using the results in Obuchowski's paper [21], we can estimate the variance of the clustered AUC,

(3)

$$\widehat{\text{var}}\left(\widehat{AUC}_c^p\right) = \frac{N_a \cdot \sum_{i=1}^{N_a}[V_a(LR_{i.}^a) - a_i \widehat{AUC}_c^p]^2}{A^2(N_a-1)} + \frac{N_u \cdot \sum_{i=1}^{N_u}[V_u(LR_{i.}^u) - u_i \widehat{AUC}_c^p]^2}{U^2(N_u-1)} + \frac{2N \cdot \sum_{i=1}^{N}[V_u(LR_{i.}^a) - a_i \widehat{AUC}_c^p]\cdot[V_u(LR_{i.}^u) - u_i \widehat{AUC}_c^p]}{AU(N-1)},$$

where

$$V_a(LR_{i.}^a) = \sum_{l=1}^{a_i}\left[\frac{1}{U}\sum_{i'=1}^{N_u}\sum_{k=1}^{u_{i'}}\varphi(LR^a(G_{il}^p), LR^u(G_{i'k}^p))\right],$$

and

$$V_u(LR_{i.}^u) = \sum_{k=1}^{u_i}\left[\frac{1}{A}\sum_{i'=1}^{N_a}\sum_{l=1}^{a_{i'}}\varphi(LR^a(G_{i'l}^p), LR^u(G_{ik}^p))\right].$$

N_a (N_u) is the total number of clusters having at least one affected (unaffected) individual. By using normal approximation, we can derive the 95% confidence interval (CI) for the clustered AUC,

(4) $$\left[\widehat{AUC}_c^p + Z_{0.025} \times \sqrt{\widehat{\text{var}}\left(\widehat{AUC}_c^p\right)}, \widehat{AUC}_c^p + Z_{0.975} \times \sqrt{\widehat{\text{var}}\left(\widehat{AUC}_c^p\right)}\right].$$

Tests containing a large number of non-causal variants tend to be unstable and less interpretable. In order to form predictive genetic tests that are more robust and interpretive, we extended the forward selection algorithm [22] to perform model selection here in family studies. The forward selection algorithm starts with a null model with no predictors, and then gradually adds predictors into the prediction model until the AUC reaches 1. In step one, we searched for a predictor, p_1, among all of the predictors, that would give the highest AUC estimate ($\widehat{AUC}_c^1$), and formed the simplest predictive genetic test, which was comprised of only one predictor, p_1. In step two, we searched for the second predictor, p_2, which, together with p_1 and with their possible interaction, reached the highest AUC estimate ($\widehat{AUC}_c^2$). While the procedure continued, the method added new predictors into the model and formed a series of models with increasing classification accuracy: $\widehat{AUC}_c^1, \widehat{AUC}_c^2, \ldots, \widehat{AUC}_c^T$. Adding more predictors makes the prediction models more complicated, so they tend to overfit the data. In order to identify the most parsimonious model with an appropriate number of predictors, we conducted a clustered K-fold cross-validation procedure. To perform the cross-validation procedure, we first divided these N families into K subsets. This could easily be done if the data contains families with similar pedigree structures (e.g. all these families are trios), but could be a challenge when families have different pedigree structures (e.g. a combination of trios and multi-generational pedigrees). The principle is to evenly assign similar types of pedigrees into different subsets and to maintain the similarity among subsets. Given these K subsets, we first used the $K-1$ subsets as the training set and applied the forward selection algorithm, constructing a series of models. The remaining subset was then used as the validation set, to estimate the predicted AUCs for each of the models built in the training set. We repeated the cross-validation process K times, with each of the K subsets serving once as the validation dataset. The predicted AUC values were averaged from models with the same number of predictors, yielding a series of averaged predicted AUC values, $\widehat{AUC}_{c,Ped}^1, \ldots, \widehat{AUC}_{c,Ped}^T$. The appropriate number of predictors, $\mathbf{p}$, was determined to be the one with the highest averaged predicted AUC, and its corresponding model was chosen as the final model. The implementation of the CORC method is written in R and will soon be available on our website: http://www.epi.msu.edu/faculty/lu/.

3. SIMULATION STUDIES

We investigated the CORC method through simulations, and compared its performance with both the GS-GEE method [7] and the forward ROC method [22] under various underlying disease models and pedigree structures. In order to implement the GS-GEE method to build a predictive genetic test, we chose the un-weighted approach to first summarize the number of risk alleles carried by each individual into a genotype score, and then fit logistic regression models with generalized estimating equations on these genotype scores. Due to a lack of powerful family-based risk prediction tools, researchers may opt to use an alternative population-based method. In this simulation, we also evaluated the performance of a population-based method, the forward ROC curve method, in a family-based risk prediction study. The forward ROC curve method shares many features with the CORC method, in that it takes interactions into account and is applicable to high-dimensional data. However, because it is a population-based method, the forward ROC curve method requires independent samples. For our simulations, we randomly selected one individual from each family and applied the forward ROC curve method.

3.1 Scenario I

We simulated three disease models, comprised of four causal loci/environmental-factors and ten non-causal loci, to investigate the performance of the CORC method, the GS-GEE method and the forward ROC method. The allele frequencies of the non-causal loci in all three of the disease models were generated randomly from a uniform distribution, ranging from 0.1 to 0.9. In the first model, four independent disease-susceptibility loci were simulated under additive modes of inheritance, with odds ratios of 1.60, 1.70,

Table 1. Comparison of the CORC method, the GS-GEE method and the forward ROC method, under three different disease models

	Simulation Setting 1			Simulation Setting 2			Simulation Setting 3		
	CORC	GS-GEE	Forward ROC	CORC	GS-GEE	Forward ROC	CORC	GS-GEE	Forward ROC
TrueAUC		0.6708			0.7337			0.6873	
Mean	0.5999	0.5925	0.5523	0.6711	0.6175	0.5943	0.6357	0.5909	0.5698
Std	0.0423	0.0358	0.0723	0.0554	0.0449	0.0988	0.0432	0.0378	0.0766
MSE	0.0068	0.0074	0.0193	0.0070	0.0155	0.0292	0.0045	0.0107	0.0197

Notes: The data were simulated under disease models in which 1) no interaction; 2) a two-way interaction; and 3) a three-way interaction existed. We summarized the mean, the standard deviation (std) and mean squared error (MSE) of the predicted AUC from each of the three methods.

1.65 and 1.55, and risk allele frequencies of 0.35, 0.35, 0.40 and 0.30, respectively. The true AUC, calculated based on these settings, was 0.6708. The second model was comprised of two independent disease-susceptibility loci and two interactive loci with their two-way multiplicative interaction [23]. The two independent loci were simulated under additive and dominant modes of inheritance, with odds ratios of 1.70 and 2.5, and risk allele frequencies of 0.35 and 0.25, respectively. The two interactive loci were assumed to follow the additive and recessive modes of inheritance, with marginal effects of 1.6 and 2.0, and risk allele frequencies of 0.4 and 0.3, respectively. To introduce the interaction, we assumed that individuals carrying risk alleles in the two interactive loci had a 2.5 times higher risk than the remaining individuals. The true AUC value for this model is 0.7337. In the third model, we introduced three disease-susceptibility loci and one two-level environmental risk factor, with odds ratios of 1.5, 2.2, 1.8 and 1.6, and risk allele frequencies/exposure frequencies of 0.4, 0.3, 0.3 and 0.45, respectively. The three disease-susceptibility loci followed additive, recessive and dominant modes of inheritance. We also introduced a three-way multiplicative interaction among the environmental risk factor, the second and third loci, by assuming that individuals carrying the risk alleles of these two loci and exposed to the environmental risk factor had a 2.2 times higher risk than the remaining individuals. The true AUC calculated based on this setting is 0.6873.

For each disease model, 1,000 replicates were generated, each consisting of 400 nuclear families and 400 sib-ships (Figure 1b). The nuclear family consisted of two parents and two offspring, while the sib-ship consisted of four offspring without parents. In each replicate, we split the whole sample into a training set and a validation set, with a ratio of 2:1. We applied all three methods to the training set to build a predictive genetic test, and then applied these tests to the evaluation set to calculate the predicted AUCs.

The results from simulation I are summarized in Table 1. In the first model, when the four disease-susceptibility loci were assumed to be independent and had similar effect sizes, our proposed CORC method attained a similar classification accuracy to the GS-GEE method. Nevertheless, when the loci were associated with different effect-sizes and interacted in the second and third models, the CORC method performed better than the GS-GEE method. For instance, when a two-way interaction was presented in the second model, the CORC method attained a predicted AUC of 0.6711, an 8.68% accuracy increase over that of GS-GEE (AUC=0.6175), and obtained an MSE of 0.0070, a 54.84% improvement over that of GS-GEE (0.0155). With a three-way interaction in the third model, the CORC method obtained an MSE of 0.0045, a 57.94% improvement over that of GS-GEE (0.0107). Due to the dramatic reduction in sample size, the forward ROC method attained the lowest performance, both in terms of the mean and MSE of the AUC estimate, in all these three models.

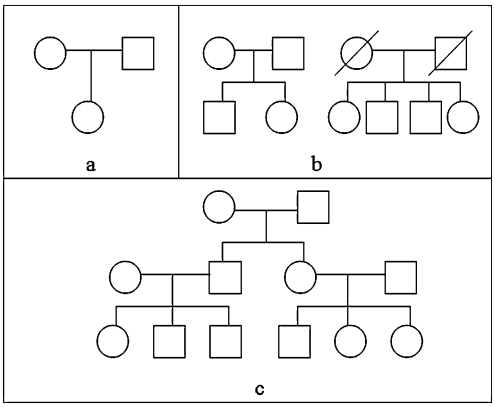

Figure 1. The three pedigree settings in simulation scenario II: **a.** *trios;* **b.** *a combination of nuclear families and sib-ships;* **c.** *the three-generation pedigrees. The units with slashes are missing individuals. The affection status of each individual was simulated according to that individual's genotypes.*

3.2 Scenario II

In simulation scenario II, we evaluated the performance of the CORC method, the GS-GEE method and the forward ROC method under three different pedigree settings. As we had done with the second disease models of simulation I,

Table 2. Comparison of the CORC method, the GS-GEE method and the forward ROC method, under three different pedigree structures

Pedigree Settings	Trios			Simple Pedigrees			Complex Pedigrees		
Test	CORC	GS-GEE	Forward ROC	CORC	GS-GEE	Forward ROC	CORC	GS-GEE	Forward ROC
Pedigree Size	1,000 Trios with 3 individuals each			400 nuclears plus 400 sib-ships with 4 individuals each			300 3-generation pedigrees with 12 individuals each		
Mean	0.6663	0.6165	0.5999	0.6711	0.6175	0.5943	0.6779	0.6187	0.5505
Std	0.0553	0.0459	0.0878	0.0554	0.0449	0.0988	0.0484	0.0424	0.1490
MSE	0.0076	0.0158	0.0256	0.0070	0.0155	0.0292	0.0054	0.0150	0.0557

Notes: The pedigree settings are illustrated in Figure 1. In this table, we have summarized the mean, the standard deviation (std) and mean square errors (MSE) of the predicted AUC from all three methods.

we simulated two independent disease-susceptibility loci and two interaction loci. Among the three pedigree settings, the first one used the same pedigree structure as in scenario I, and the remaining two were made up of 1,000 trios and 300 three-generation pedigrees (Figure 1). The total sample sizes for the three pedigree settings were 3,000, 3,200 and 3,600, respectively, and were comparable overall.

The results from simulation scenario II are summarized in Table 2. In all three of the pedigree settings, tests built by the proposed CORC method tended to be more robust and accurate than those built by the GS-GEE method. For instance, in the setting with three-generation pedigrees, the CORC method reached a predicted AUC of 0.6779, a 9.57% increase over that of GS-GEE (0.6187). It also attained an MSE of 0.0054, a 64% improvement over that of GS-GEE (0.0150). As the pedigree structure became more complex, tests built from both CORC and GS-GEE obtained a slight increase in classification accuracy. However, the CORC method tended to gain a greater increase than GS-GEE. For instance, when the pedigree structure changed from trios to a combination of nuclear families and sib-ships, the predicted AUC mean of CORC attained an increase of 0.72%, from 0.6663 to 0.6711, which was greater than that of GS-GEE (a 0.16% increase, from 0.6165 to 0.6175). Similar to simulation I, the forward ROC method had the lowest performance in all three settings. When the pedigree structure became more complex, the test built by the forward ROC method tended to be less accurate, with a predicted AUC value decrease of 8.97%, from 0.5999 to 0.5505. This decrease in classification accuracy can be explained by the decreased sample size, from 1,000 in the trios setting to 300 in the three-generation setting.

4. DATA APPLICATION FOR CONDUCT DISORDER

Conduct disorder (CD) is a disorder of childhood and adolescence that involves chronic behavior problems, such as defiant, impulsive, antisocial behavior, drug use and criminal activity. Using the proposed method and the two existing methods described earlier, we analyzed the large scale International Multi-Center ADHD Genetics Project (IMAGE) dataset to evaluate predictive genetic tests for CD. The IMAGE project, as part of the Genetics Analysis Information Network (GAIN) initiative [24], is one of the largest GWAS conducted to date, and is designed to investigate the genetic causes of Attention-deficit/hyperactivity disorder (ADHD) and CD. The IMAGE study contains over nine hundred parent-child trios, which were genotyped using the Perlegen 600K SNP platform.

In this study, we assessed the combined effect of gender [25] and 46 CD-associated loci from recent GWAS [24,26] on predicting CD. Similarly, we split the entire CD dataset into a training set and a validation set, with a 2:1 ratio. The training set was used to build a predictive genetic test, and the validation set was used to estimate the predicted classification accuracy of the test. Figure 2 plotted the ROC curves of the CD test formed by these three methods. Consistent with the simulation results, the CORC method attained the highest classification accuracy among these three methods. In the validation set, the CORC method obtained a predicted AUC value of 0.6908, 19.14% higher than that of GS-GEE (a predicted AUC of 0.5798) and 0.73% higher than that of the forward ROC method (a predicted AUC of 0.6858). We also calculated the 95% confidence intervals (CIs) of the predicted AUC. The AUC estimate from the CORC method had greater precision (CI: 0.6332-0.7484) than those from the GS-GEE (CI: 0.5163-0.6433) and the forward ROC methods (CI: 0.5860-0.7856). In this application, the GS-GEE method had the lowest performance, perhaps due to the violation of equal effect-sizes and independence assumptions. Although the forward ROC method attained higher classification accuracy than GS-GEE, we noted that the forward ROC method obtained a wider CI than both GS-GEE and CORC, mainly due to the reduced sample size.

The predictive test built by the CORC method selected gender, *rs10492664*, *rs10797919* and *rs1644305* as predictors in the final model. Among these, *rs10492664* is one

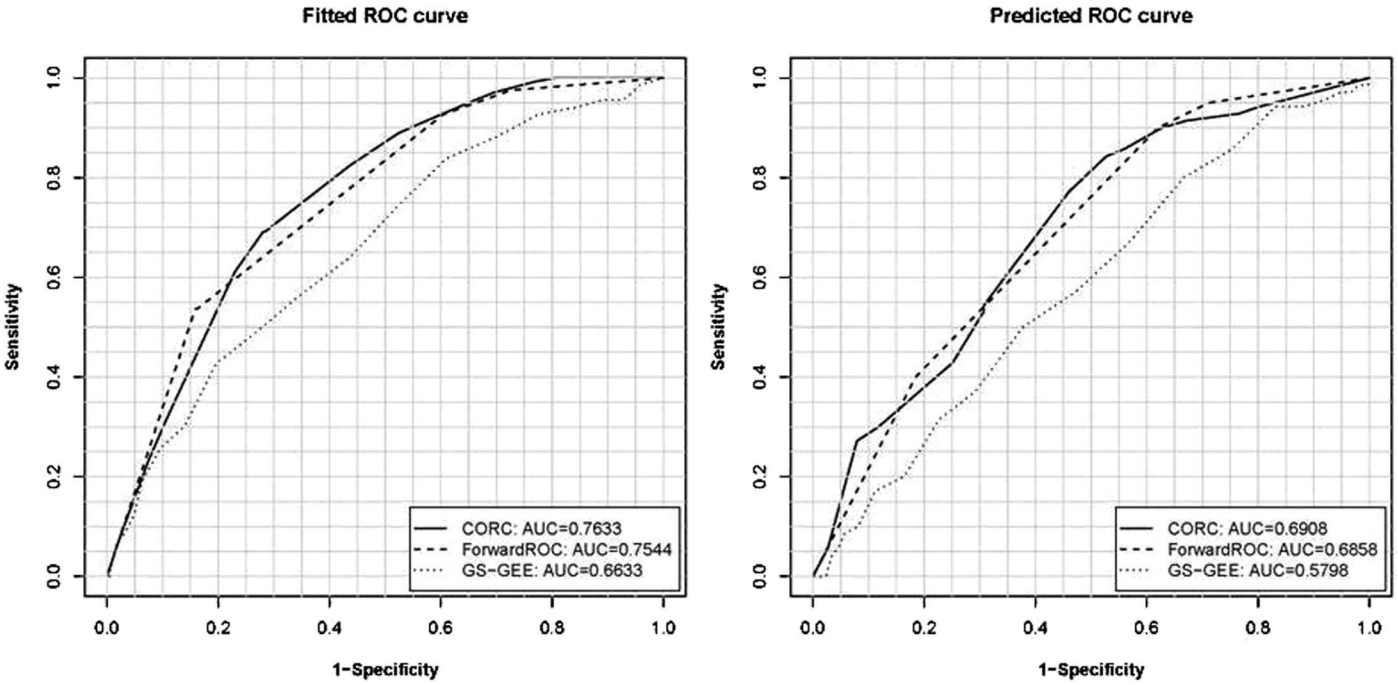

Figure 2. The ROC curves of the CD test formed by the three methods. The left panel shows the ROC curves of the CD test built on the training set. The ROC curves in the right panel were plotted by applying the built CD test to the validation set.

of the top five association signals in the previous GWAS study [24] and *rs10797919* is in a chromosomal region previously found to show linkage with drive-for-thinness and obsessionality. Similar to CD, this disease can be regarded as a dysregulation of serotonergic neurotransmission [27].

5. DISCUSSION

In this study, we introduced a clustered optimal ROC curve method (CORC) for family-based risk prediction studies. The CORC method incorporates a clustered ROC curve algorithm with no restriction on pedigree structures, and, thus, can be applied to data with complex pedigree structures or a mixture of pedigree structures. As shown in the simulation scenario II, the CORC method performs robustly in different settings of pedigree structures. It has the same advantages as the forward ROC method, such as having a theoretically optimal performance and considering interactions. However, as illustrated by our simulations and the real data application, population-based risk prediction methods, such as the forward ROC method, are subject to low performance when there is family data, because they require independent samples.

An existing method for family-based risk prediction research, the GS-GEE method, requires assumptions—such as independence among testing variants—and could be subject to low performance if these assumptions fail. As illustrated by simulation I, when the disease-susceptibility loci were independently associated with the disease outcome, the proposed CORC method and the GS-GEE method performed similarly. However, as revealed in both the simulations and the real data application, the classification accuracy of the tests from the GS-GEE method was much lower than that from the CORC method when interactions were presented. Thus, the proposed CORC method is more suitable for the risk prediction of complex diseases, which likely involves multiple risk factors and interactions.

In a recent study, Wray et al. revealed that the maximum AUC value of a test was constrained by the disease's prevalence and heritability [28]. Given the estimated heritability (∼50%) and the cumulative prevalence (∼9%) derived from previous CD studies, we estimated that the AUC value of CD could maximally reach 0.88 [28,29,30]. Recall that the CD test from our study attained a predicted AUC of 0.6908, which is lower than this estimated maximum AUC, and is unlikely to be immediately useful in clinical practice. This could be due to our study's use of only a handful of loci with significant marginal effects. Loci with small or even no marginal effects may play important roles in disease pathways, and likely interact with other genetic variants to provide predictive values. Note that the proposed method, which is similar to the forward ROC curve method, can be applied to high-dimensional data. Following this initial study, one of the natural next steps is to conduct a risk prediction study on a much larger number of risk factors, including loci with small, or even no, marginal effects for disease prediction.

When constructing prediction models based on high-dimensional data, a large amount of computation time

will be required. In our previous study, we conducted a population-based genome-wide risk prediction analysis on a 500K Wellcome Trust rheumatoid arthritis GWAS dataset, comprised of approximately 5,000 individuals. The whole genome-wide risk prediction analysis, including the cross-validation process, took 43 hours on a high performance workstation by using a C++ program we wrote. This C++ program can be modified for a family-based risk prediction analysis. For a dataset with the same sample size, we anticipate that a similar amount of time (i.e., approximately two days) would be required for the whole genome-wide family-based risk prediction of a 500K dataset.

The clinical utility of a predictive genetic test depends on the nature of the disease, the availability of a prevention method, and the cost of screening and surveillance measures [31]. It has been shown elsewhere that some specific combinations of genetic variants have failed to make any significant improvements in predicting cardiovascular disease and multiple sclerosis [13,32]. Generally, for diseases with a modest prevalence and sibling recurrence risk (λ_s), genetic prediction has less value for clinical practice. However, it could still be of great value in making treatment decisions. For some specific symptoms of CD, pharmacotherapy (e.g., stimulants, anti-depressants, lithium and anticonvulsants) has become an adjunct treatment for this disease [33]. Thus, we expect that genetic tests predicting a subgroup of patients who respond to a particular drug could have great value for the development of more effective personalized prevention and treatment methods.

ACKNOWLEDGMENTS

We thank Karen Friderici and Yuehua Cui for many helpful comments and suggestions on the manuscript. We also appreciate critical input from two anonymous reviewers. The data sets used for the application analysis described in this manuscript were obtained from the Genetic Association Information Network (GAIN) Database, found at http://www.ncbi.nlm.nih.gov/gap through the dbGAP accession number phs000016.v2.p2.

Received 31 May 2010

REFERENCES

[1] Ginsburg, G. S. and Willard, H. F. (2009). Genomic and personalized medicine: foundations and applications. *Transl. Res.* **154** 277–287.

[2] Evans, J. P., Skrzynia, C. and Burke, W. (2001). The complexities of predictive genetic testing. *BMJ.* **322** 1052–1056.

[3] Epstein, C. J. (2006). Medical genetics in the genomic medicine of the 21st century. *Am. J. Hum. Genet.* **79** 434–438.

[4] Taylor, S. D. (2004). Predictive genetic test decisions for Huntington's disease: context, appraisal and new moral imperatives. *Soc. Sci. Med.* **58** 137–149.

[5] Ioannidis, J. P. (2009). Prediction of cardiovascular disease outcomes and established cardiovascular risk factors by genome-wide association markers. *Circ. Cardiovasc. Genet.* **2** 7–15.

[6] Johansen, C. T. and Hegele, R. A. (2009). Predictive genetic testing for coronary artery disease. *Crit Rev. Clin. Lab Sci.* **46** 343–360.

[7] Meigs, J. B. et al. (2008). Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* **359** 2208–2219.

[8] Lango, H. et al. (2008). Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes.* **57** 3129–3135.

[9] Lu, Q. and Elston, R. C. (2008). Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. *Am. J. Hum. Genet.* **82** 641–651.

[10] 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.

[11] van Hoek, M. et al. (2008). Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. *Diabetes.* **57** 3122–3128.

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

[13] Paynter, N. P. et al. (2010). Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA.* **303** 631–637.

[14] Meuwissen, T. H., Hayes, B. J. and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics.* **157** 1819–1829.

[15] de los Campos, G. et al. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics.* **182** 375–385.

[16] Baker, S. G. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics.* **56** 1082–1087. MR1815586

[17] Lu, Q. et al. (2009). Using the optimal robust receiver operating characteristic (ROC) curve for predictive genetic tests. *Biometrics.*

[18] McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics.* **58** 657–664. MR1926119

[19] Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* **39** 561–577.

[20] Egan, J. P. (1975). Signal Detection Theory and ROC Analysis. New York, Academic Press.

[21] Obuchowski, N. A. (1997). Nonparametric analysis of clustered ROC curve data. *Biometrics.* **53** 567–578.

[22] Ye, C. Y. et al. (2010). A non-parametric method for building predictive genetic tests on high-dimensional data, with an application to rheumatoid arthritis. *Submitted.*

[23] 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.

[24] Anney, R. J. et al. (2008). Conduct disorder and ADHD: evaluation of conduct problems as a categorical and quantitative trait in the international multicentre ADHD genetics study. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B** 1369–1378.

[25] Zahn-Waxler, C., Shirtcliff, E. A. and Marceau, K. (2008). Disorders of childhood and adolescence: gender and psychopathology. *Annu. Rev. Clin. Psychol.* **4** 275–303.

[26] Sonuga-Barke, E. J. et al. (2008). Does parental expressed emotion moderate genetic effects in ADHD? An exploration using a genome wide association scan. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B** 1359–1368.

[27] Devlin, B. et al. (2002). Linkage analysis of anorexia nervosa incorporating behavioral covariates. *Hum. Mol. Genet.* **11** 689–696.
[28] Wray, N. R. et al. (2010). The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS. Genet.* **6** e1000864.
[29] Dick, D. M. et al. (2010). Genome-wide association study of conduct disorder symptomatology. *Mol. Psychiatry.*
[30] Costello, E. J. et al. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Arch. Gen. Psychiatry.* **60** 837–844.
[31] Evans, J. P., Skrzynia, C. and Burke, W. (2001). The complexities of predictive genetic testing. *BMJ.* **322** 1052–1056.
[32] Sawcer, S. et al. (2010). What role for genetics in the prediction of multiple sclerosis? *Ann. Neurol.* **67** 3–10.
[33] Campbell, M., Gonzalez, N. M. and Silva, R. R. (1992). The pharmacologic treatment of conduct disorders and rage outbursts. *Psychiatr. Clin. North Am.* **15** 69–85.

Chengyin Ye
Department of Bioinformatics
College of Life Science
Zhejiang University
Hangzhou, Zhejiang 310058
P.R. China

Department of Epidemiology
Michigan State University
East Lansing, Michigan 48824
USA
E-mail address: yechengyin@zju.edu.cn; cye@epi.msu.edu

Jun Zhu
Institute of Bioinformatics
Zhejiang University
Hangzhou, Zhejiang 310029
P.R. China
E-mail address: jzhu@zju.edu.cn

Qing Lu
Department of Epidemiology
Michigan State University
East Lansing, Michigan 48824
USA
E-mail address: qlu@epi.msu.edu