Published in final edited form as:

J Biopharm Stat. 2010 March ; 20(2): . doi:10.1080/10543400903572811.

BAGGING OPTIMAL ROC CURVE METHOD FOR PREDICTIVE GENETIC TESTS, WITH AN APPLICATION FOR RHEUMATOID ARTHRITIS

Qing Lu¹, Yuehua Cui², Chengyin Ye¹, Changshuai Wei¹, and Robert C. Elston³

¹Department of Epidemiology, Michigan State University, East Lansing, Michigan

²Department of Statistics and Probability, Michigan State University, East Lansing, Michigan

³Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA

Abstract

Translation studies have been initiated to assess the combined effect of genetic loci from recently accomplished genome-wide association studies and the existing risk factors for early disease prediction. We propose a bagging optimal receiver operating characteristic (ROC) curve method to facilitate this research. Through simulation and real data application, we compared the new method with the commonly used allele counting method and logistic regression, and found that the new method yields a better performance. The new method was applied on the Wellcome Trust data set to form a predictive genetic test for rheumatoid arthritis. The formed test reached an area under the curve (AUC) value of 0.7.

Keywords

Area under the ROC curve; Bootstrap aggregation; Gene-gene interaction; Genomewide association studies

1. INTRODUCTION

The success of recently accomplished genome-wide association studies (GWAS) provides a unique opportunity for early disease prediction and prevention (Guttmacher and Collins, 2005; Khoury et al., 2007). The findings from GWAS have improved our understanding of disease biology pathways, and this may eventually promote effective new prevention or treatment methods. But what is probably more important and attainable at this time is the use of the novel loci discovered through GWAS, as well as other established risk factors, to predict an individual's risk of developing diseases (i.e., to form a predictive genetic test).

Copyright © Taylor & Francis Group, LLC

Address correspondence to Qing Lu, PhD, Department of Epidemiology, Michigan State University, B601 West Fee Hall, East Lansing, Michigan 48824, USA; qlu@epi.msu.edu.

Publisher's Disclaimer: Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Predictive genetic tests can identify high-risk populations early in life so that the current available prevention strategies—such as lifestyle intervention—can be applied early enough to effectively prevent individuals from contracting the disease (Bell, 1998; Epstein, 2006; Evans et al., 2001).

Recent studies have been conducted to investigate the role of currently identified genetic and clinical risk factors, in particular those from recent GWAS, in early disease prediction (Lango et al., 2008; Meigs et al., 2008; van Hoek et al., 2008; Zheng et al., 2008). Receiver operating characteristic (ROC) curve analysis was commonly adopted in these studies to evaluate the performance of the predictive genetic test (Janssens et al., 2004). The ROC curve plots a test's sensitivity against its specificity by continuously changing the cutoff points over the whole range of possible test results, and is often summarized by its one-dimension summary index—the area under the ROC curve (AUC) (McClish, 1989).

Most of these studies were interested in predicting common complex diseases that are caused by multiple genetic and environmental risk factors. Various methods have been used in these studies to combine multiple predictors for the ROC curve analysis. Among those, the most commonly adopted is the allele counting method (Meigs et al., 2008; Zheng et al., 2008). Given multiple genetic predictors, the allele counting method derives a person's risk score by simply counting the number of risk allele that he/she carries across all risk loci. For instance, if an individual has homozygous risk genotypes at all p loci, he/she will have a risk score of $2p$. By calculating such a score for all individuals, multiple risk factors are summarized into a one-dimensional risk score. Existing software (e.g., *R*) can then be implemented to plot the ROC curve and estimate the area under the curve (AUC). The advantage of this method is that it is straightforward and easy to implement. However, the method has certain limitations, as mentioned in several studies (Lango et al., 2008; van Hoek et al., 2008). It is based on an assumption that each risk allele has the same predisposing risk for disease, and that the risks are additive within and across loci. This might not hold when a few loci have stronger effects, or when interaction exists. If the assumption fails, then the allele counting method could yield a low performance (van Hoek et al., 2008).

Another commonly used method is logistic regression (Lango et al., 2008; Lu et al., 2009b; van Hoek et al., 2008). When dealing with multiple genetic predictors, logistic regression assumes a linear relationship among the predictors and uses a logit link to combine them into a one-dimensional fitted value. Given this one-dimensional value, the ROC curve can be formed. Recent studies, however, have shown that logistic regression is not always optimal for ROC curve analysis, despite its common usage (Pepe et al., 2006). Furthermore, logistic regression is a parametric method, so it may not be suitable for common complex diseases caused by a large number of risk factors and their interactions. For instance, recent studies identified more than 30 loci predisposing to Crohn's disease (Barrett et al., 2008). A logistic regression model that considers 30 genetic predictors and their two-way interactions will require 465 parameters. When the sample size is limited, fitting a model with such a large number of parameters will lead to a biased estimate (Peduzzi et al., 1996).

Considering these issues, this study proposes a bagging optimal ROC curve method. The new method is developed based on the theory of the optimal ROC curve, which forms an ideal ROC curve using the likelihood ratio rule (Egan, 1975). It adopts a bootstrap aggregation approach (Breiman, 1996), and therefore can have robust performance across different samples. The new method is nonparametric, and therefore does not suffer from the issue of an increasing number of parameters. We compare the new method with the allele counting method and logistic regression using simulation, and find that the new method performs better. Finally, we apply the three methods to the Wellcome Trust Rheumatoid

Arthritis GWAS data set, and form a new predictive genetic test for rheumatoid arthritis (RA). The RA test formed by the bagging optimal ROC curve attains the highest classification accuracy (AUC = 0.7043).

2. METHOD

2.1. Optimal ROC Curve and Its Use for Multiple Predictors

Sensitivity and specificity are commonly used as a pair to measure a test's accuracy. Given the disease status (D denotes disease and $\bar{D}$ denotes non-disease) and test result (x denotes a positive test result and $\bar{x}$ denotes a negative test result), sensitivity (Se) is defined as the probability of a positive result, given that the disease is present, $Se = P(x|D)$, and specificity (Sp) is defined as the probability of a negative test result, given that the disease is absent, $Sp = P(\bar{x}|\bar{D})$. When there are multiple pairs of sensitivity and specificity values available, the test's accuracy is evaluated by the ROC curve (Zweig and Campbell, 1993). The ROC curve consists of the entire set of sensitivity and specificity pairs resulting from the continually changing cutoff over the range of test results, and therefore has been recognized as a global measure of a test's accuracy (McClish, 1989).

The likelihood ratio (LR) corresponds to the slope of the ROC curve, and can be expressed as a simple function of sensitivity and specificity,

$$LR(x) = \frac{Se}{Sp} = \frac{P(x|D)}{P(\bar{x}|\bar{D})}$$

The LR itself is not a good measure of a test's classification accuracy (Pepe, 2003). However, it plays an important role in forming the optimal ROC curve. In the 1970s, Egan (1975) found an ideal ROC curve—named the proper ROC curve or optimal ROC curve—that can be formed by ranking the test's results according to their LR values. It has been shown that each point on the optimal ROC curve: (1) attains the highest sensitivity for any fixed value of specificity; (2) has the highest specificity for any fixed value of sensitivity; and (3) has the smallest misclassification probability and expected cost (Egan, 1975; McIntosh and Pepe, 2002). The optimal ROC curve also achieves the highest classification accuracy, in terms of the AUC (Lu and Elston, 2008).

The optimal ROC curve was originally introduced for a single predictor, but its usefulness for combining multiple predictors was recently discovered. Based on the concept of the optimal ROC curve, Baker proposed three methods for combining multiple cancer biomarkers for disease prediction and showed, in a real application, that the optimal ROC curve methods performed better than logistic regression (Baker, 2000). More recently, we also introduced two optimal ROC curve-based methods to combine multiple single-nucleotide polymorphism (SNP) markers. For common complex diseases, where the underlying disease model is unknown, we found that only one optimal ROC curve-based method could obtain an unbiased AUC estimate, while the other four methods tend to overestimate a test's discriminative ability (Lu et al., 2009a).

The method uses a backward clustering algorithm to select a model that most represents the underlying disease model, and so obtains an unbiased AUC estimate. Assuming we have p loci, the backward clustering algorithm gradually groups the p -locus genotypes. By doing that, it generates a series of models, from the most complex one, which treats each p -locus genotype as a separate cluster, to the simplest one that clusters all p -locus genotypes into one group. Cross-validation is then used to choose the best model, with the appropriate model complexity. By clustering the p -locus genotypes, the “noise” loci are automatically

eliminated and the markers' modes of inheritance are approached. Thus the algorithm provides a robust performance under the common complex disease scenario. The backward clustering algorithm could also take the interaction effect into account. For instance, assume that we have two disease-susceptibility loci, A and B , and an interaction between two loci that follows a threshold model (Marchini et al., 2005). Each locus has three genotypes, so we could have nine possible multi-locus genotypes in total. Ideally, by continuously grouping these nine multi-locus genotypes, the backward clustering algorithm eventually clusters them into two groups: A high-risk group consists of multi-locus genotypes having at least one of the disease-susceptibility alleles; a low risk group has the remaining multi-locus genotypes. By ranking the two risk groups in descending order of their LRs, and plotting the optimal ROC curve, we incorporate the interaction effect between two loci into the prediction model. Although a simple example with two-locus interaction is illustrated here, the same argument can be applied to more complex scenarios, such as scenarios with a higher order gene-gene/gene-environment interaction and scenarios with more than one gene-gene/gene-environment interaction.

2.2. Bagging Optimal ROC Curve Method

We adopt a bootstrap aggregation procedure in our method. A model is built by implementing the backward clustering algorithm on each bootstrap replicate drawn from part of the data, and is then applied to the remaining sample to form the optimal ROC curve. The bagging optimal ROC curve is obtained by averaging the optimal ROC curves from all bootstrap replicates.

We illustrate the method with p genetic predictors. Given the data, we randomly divide it into two groups—one for the purpose of training and the other for the purpose of evaluation. In total, there are B bootstrap replicates drawn from the training data set, and the backward clustering algorithm is applied to each bootstrap replicate to search for the best model.

Assuming $G^{(b)} = (g_1^{(b)}, g_2^{(b)}, \dots, g_k^{(b)})$ is the set of multi-locus genotype clusters obtained from the backward clustering algorithm on the b th bootstrap replicate, we estimate their LRs by

$$LR(g_j^{(b)}) = \frac{p(g_j^{(b)} | D)}{p(g_j^{(b)} | \bar{D})} \quad j = 1, 2, \dots, K.$$

We apply the selected model to the validation data set, to cluster the multi-locus genotype, and then to rank the resulting multi-locus genotype clusters by their corresponding LRs from the bootstrap replicate. By doing that, each individual in the validation data set is assigned a rank in such a way that individuals carrying high-risk multi-locus genotypes (i.e., large LRs) rank first and individuals carrying low-risk multi-locus genotypes (i.e., small LRs) rank last. Since the ROC curve depends only on the rank, we use the ranks of individuals in the validation data set $O_1^{(b)} = (o_1^{(b)}, o_2^{(b)}, \dots, o_n^{(b)})$, to represent the optimal ROC curve formed on the b th bootstrap replicate. We repeat this for all B bootstrap replicates. The bagging optimal ROC curve is obtained by averaging the ranks from the B bootstrap samples, $O^{(avg)} = (O_1^{(avg)}, O_1^{(avg)}, \dots, O_n^{(avg)})$, where

$$O_1^{(avg)} = \frac{1}{B} \sum_{b=1}^B o_1^{(b)}.$$

The bagging optimal ROC curve consists of a set of sensitivity and specificity pairs that result from the continually changing cutoffs, t , over the range of the average ranks,

$$\begin{cases} Se_t = \frac{\sum_{i=1}^{N_D} I(o_i^{(avg)} \leq t)}{N_D} \\ Sp_t = \frac{\sum_{j=1}^{N_{\bar{D}}} I(o_j^{(avg)} > t)}{N_{\bar{D}}} \end{cases} \quad (1)$$

where $t = [\min(O^{(avg)}), \max(O^{(avg)})]$. N_D and $N_{\bar{D}}$ are the number of cases and controls in the validation data set, respectively.

The area under the optimal ROC curve can be calculated by

$$AUC^{(b)} = \frac{1}{n_D n_{\bar{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} \psi[O_i^{(avg)}, O_j^{(avg)}] \quad (2)$$

where the kernel function, $\psi[o_i^{(avg)}, o_j^{(avg)}]$ is given by

$$\psi[o_i^{(avg)}, o_j^{(avg)}] = \begin{cases} 1 & \text{if } o_i^{(avg)} < o_j^{(avg)} \\ 0.5 & \text{if } o_i^{(avg)} = o_j^{(avg)} \\ 0 & \text{if } o_i^{(avg)} > o_j^{(avg)} \end{cases}$$

3. RESULTS

3.1. A Predictive Genetic Test for RA Using the Wellcome Trust GWAS Data

The recent intensive genetic research for RA has identified numerous novel RA loci (Gorman et al., 2004; Lee et al., 2005; Plenge et al., 2005, 2007a,b; Thomson et al., 2007; Vignal et al., 2009; Zervou et al., 2008). This study considered gender and 13 well-replicated RA loci (Table 1), and assessed their combined roles in predicting RA, using the Wellcome Trust RA GWAS data set.

The Wellcome Trust GWAS (Wellcome Trust Case Control Consortium, 2007) is one of the largest and most comprehensive case-control studies to date, aimed at discovering genetic contributions to seven common complex diseases, including RA. The RA cases in that study were recruited from different regions across the United Kingdom. One-half of the controls were chosen from the 1958 British Birth Cohort, and the remaining half came from UK Blood Service controls. The data have been reformatted and cleaned for our analysis. Non-Caucasian individuals were removed from the study, due to their small numbers ($n = 64$). Records of low quality (e.g., resulting from low DNA quality) were also excluded. The final sample for analysis was comprised of 1860 RA cases and 2938 controls. A subset of data was then created by extracting gender and the genotypes of 13 loci from the GWAS data set. For the genetic loci, we directly obtained two loci from the GWAS data set and imputed the remaining 11 loci that were not present in the data set by using the model-based imputation software IMPUTE (Marchini et al., 2007). This software has been applied to the Wellcome Trust 1958 British Birth Cohort control data set, yielding an accuracy rate of 98.4%. We applied the IMPUTE software to infer the 11 loci, and summarized the imputation result in Table 1. The average maximum posterior genotype call probability for each locus is greater than 0.98, indicating high imputation accuracy. We applied the bagging optimal ROC curve method, logistic regression, and the allele counting method on the created subset to form a predictive genetic test for RA. Each method combined gender and 13 genetic predictors into a one-dimensional risk score. We varied the cutoff over the range of the one-dimensional risk score in order to both plot the ROC curve and estimate the AUC. The AUC was then used as a measure for evaluating the test's discriminative ability.

Since the model of inheritance for each locus is unknown, a model used in the previous type 2 diabetes (T2D) predictive genetic test study (Lango et al., 2008) was adopted for the logistic regression analysis. The model considers only single locus effects, and codes each locus as two dummy variables, x_{1l} and x_{2l} ,

$$\text{logit } \mu = \alpha + \gamma z + \sum_{i=1}^p (\beta_{i1} x_{i1} + \beta_{i2} x_{i2})$$

where

$$Z = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}, x_{i1} = \begin{cases} 1 & \text{if } AA \\ 0 & \text{if } Aa/aa \end{cases} \text{ and } x_{i2} = \begin{cases} 1 & \text{if } Aa \\ 0 & \text{if } AA/aa \end{cases}$$

Because building models and estimating the AUC based on the same data set will lead to an overestimated AUC, we randomly divided the data into two parts, using two-thirds of the data to build the model, and then applying the model to the remaining one-third of the data set to plot the ROC curve and estimate the AUC.

The logistic regression model just described does not take any interaction effect into account, and may result in low performance if there is a strong interaction effect among gender and the 13 loci. Therefore, in addition to conducting the logistic regression analysis that only considered a single-locus effect, we also fitted a logistic regression model that included all possible two-way interactions.

The bagging optimal ROC curve method introduced in the method section was also applied to the data. We used the same training data sets and validation data sets as were used in the logistic regression analyses, and drew 100 bootstrap replicates from the training data set. The models were first built on the 100 bootstrap samples and were then applied to the validation data sets to plot the ROC curve and estimate the AUC.

For the allele counting method, we summarized the risk alleles at the 13 loci, to generate a risk score for each individual. Since females have a higher risk of RA, we increased the risk score by 1 if the individual was female. The allele counting method does not require a model-building step, so we used the entire data set to plot the ROC curve and estimate the AUC.

Figure 1 plots the ROC curves from the four models. The predictive genetic tests formed by the bagging optimal ROC curve method attained the highest classification accuracy (AUC = 0.7043). We found that logistic regression that only looks at single locus effects has a performance comparable to that of the bagging optimal ROC curve method (AUC = 0.7037). However, logistic regression that included both single-locus and interaction effects seriously underestimated the test performance (AUC = 0.5564). This low performance is mainly due to an overly complex model built from the training data set with extra interaction terms. As compared to the bagging optimal ROC curve method and logistic regression with single locus effects, the test from the allele counting method also achieved a low discriminative ability, with an AUC value of 0.6738.

It should be noted that, in general, the SNPs selected for the predictive genetic test need to come from independent studies. Otherwise, the prediction model may be overfitted. In our study, two loci, rs6457617 and rs6920220, were chosen from the Wellcome Trust RA GWAS. However, in view of the fact that these two loci were also found to be associated with RA in other studies (Han et al., 2009; Thomson et al., 2007), overfitting is less of a concern here.

3.2. Simulations

We further evaluated the performance of the proposed method, logistic regression and the allele counting method using simulations. The simulation studies were conducted in two disease scenarios—one without gene–gene interaction and one with gene–gene interaction.

3.2.1. A Simple Disease Model Without Gene–Gene Interaction—We first simulated data under a simple disease scenario using three disease loci and four noncausal loci, assuming no interaction among the three disease loci. The three disease loci were assumed to have a small/medium effect size, with odds ratios of 1.5, 1.2, and 1.6, respectively. We simulated the three loci under recessive, additive, and dominant modes of inheritance, with disease allele frequencies of 0.3, 0.2, and 0.4, respectively. The allele frequencies of the four noncausal loci were randomly sampled from a uniform distribution; we assumed that these four loci have no effect on the disease.

One thousand replicates were simulated under each of two sample size situations ($n = 500$, 1000). Similar to the real data application, we randomly divided each replicate into training and validation data sets. We applied the proposed method and logistic regression on the training data to construct the models, and then fitted the models onto the validation data sets to estimate the AUCs. Since the allele counting method does not require an independent data for model building, we used the entire data set to estimate the AUC.

In all scenarios, the bagging optimal ROC curve method performed better than logistic regression or the allele counting method, in terms of having higher AUC values with smaller standard deviations (Table 2). Logistic regression, particularly one with the interaction terms, tends to underestimate AUC values. This could be due to the overfitted models built in the training data sets, with their extra parameters for the noncausal loci, and/or the interaction terms. The allele counting method also produced lower AUC estimates. Since the allele counting method assumes similar effect sizes for each locus, and in this simulation the effect sizes for the loci were different (i.e., the highest odds ratio for a disease locus was 1.6 and odds ratios for the noncausal loci were 1), this led to the method's low performance.

3.2.2. A Disease Model with Higher Order Gene–Gene Interactions—Gene–gene interactions may play an important role in causing common complex diseases. Incorporating these gene–gene interactions into the test could further improve the test's accuracy. For this study, we simulated a disease model with a high order of gene–gene interaction, and evaluated the performance of the different methods in the presence of this gene–gene interaction. We adopted the same simulation setting as already described, but introduced an interaction among the three disease loci by simply doubling the risk for individuals carrying at least one disease susceptibility allele at all three loci (Marchini et al., 2005).

Again, the bagging optimal ROC curve method achieved better AUC estimates, with higher AUC values and smaller standard deviations (Table 3). As expected, an increased sample size improved estimates from all three methods, with increasing AUC values and smaller standard deviations. Compared with the results from the previous simulation, the estimates were subject to higher standard deviations. This may be due to the increasing complexity of the underlying disease model, which resulted from introducing an additional gene–gene interaction.

Although the simulation conducted here is based on a specific interaction model, the threshold model (Marchini et al., 2005), we anticipate the same conclusion will hold for most of the other interaction models. For instance, in the case of low or no marginal effects but high interactions, the test formed either by the allele counting method or by logistic regression that models only the marginal effect will obtain lower classification accuracy

because neither of the two methods counts the interaction effect. Logistic regression with the interaction term is also subject to lower accuracy since the low performance of the method in our simulation is mainly due to overparameterization of the model, rather than the underlying interaction model.

4. CONCLUSIONS

Recent studies have been conducted to evaluate the role of multiple genetic and clinical risk factors in predicting common complex diseases (Lango et al., 2008; Meigs et al., 2008; van Hoek et al., 2008; Zheng et al., 2008). Various methods have been used for these studies. However, the performance of these methods has not been carefully evaluated. In this study, we proposed a bagging optimal ROC curve method, and compared its performance with the other two popularly adopted methods, the allele counting method and logistic regression.

The proposed bagging optimal ROC curve method can be looked upon as an application of the optimal ROC curve theory introduced in the early 1970s (Egan, 1975). This theory simply shows that the ROC curve that is based on the LR rule attains the highest performance at each point on the ROC curve, and in overall discriminative ability. The new method also adopts a bootstrap aggregation procedure (Breiman, 1996), which leads to a robust performance.

Through simulation and real data application, we showed that a test built by using the bagging optimal ROC curve method had higher classification accuracy than that built by logistic regression or the allele counting method. Using the same simulation model with a high order gene–gene interaction and a sample size of 1000, we also found that the new method (AUC = 0.6230, SD = 0.0148) had an improved performance over the previously proposed optimal robust ROC curve method (Lu et al., 2009a) (AUC = 0.5982, SD = 0.03496).

In recent studies (Meigs et al., 2008; Zheng et al., 2008), the allele counting method has been widely adopted for investigating a predictive genetic test for common complex diseases. The allele counting method is easy to implement and gives an interpretable result (e.g., an individual carrying the largest number of disease alleles has the highest risk for disease). However, as pointed out in several articles (Lango et al., 2008; van Hoek et al., 2008), the method makes a strong assumption that could fail in many circumstances. For instance, it assumes the same risk for each locus. This may not be valid for diseases such as type 1 diabetes and RA, in which some loci (e.g., human leukocyte antigen loci) have much stronger effects than others. It also assumes the disease alleles follow an additive model, both within and across loci. This will not hold true if the disease allele follows a dominant/recessive model, or when there is any interaction. These assumptions should be carefully evaluated before applying the method; otherwise, as shown in our study, one can expect a low performance of the method. An improved allele counting method, the weighted allele counting method (Purcell et al., 2009), has recently been introduced. It counts the number of risk alleles and weights each SNP by its log odds ratio, and thus does not assume an equal effect size across loci. Through simulation (i.e., by assuming a high order of gene–gene interaction and a sample size of 1000), we found that the weighted allele counting method (AUC=0.6081, SD = 0.0315) was improved over the allele counting method (AUC = 0.5800, SD = 0.0187). It should be noted that the weighted allele counting method still assumes an additive mode of inheritance within loci. In our simulation, the weighted allele counting method had a lower performance than the bagging optimal ROC curve method (AUC = 0.623, SD=0.0148), which may be explained by the violation of such an assumption.

Logistic regression is commonly used in the area of diagnostic tests (Pepe, 2003). It is convenient to use through existing software, such as SAS and R, and is a powerful tool for forming predictive genetic test for common complex diseases. In our real data application, logistic regression had a performance comparable to that of the bagging optimal ROC curve method. It has certain limitations, however, for developing predictive genetic tests when interactions exist. For a logistic regression model that considers the gene–gene interaction effect, the number of parameters required to model interactions increases exponentially with the increasing number of loci. This could lead to a biased estimate (Peduzzi et al., 1996). As illustrated by both simulations and real data application, a logistic regression model considering two-way gene–gene interactions could seriously underestimate the AUC values.

The completion of GWAS for common complex diseases led to the discovery of novel disease-susceptibility loci. A promising strategy will be to use these genetic findings to identify a high-risk population early, and then apply preventive methods to delay, or even prevent, the clinical onset of disease (Guttmacher and Collins, 2005; Khoury et al., 2007). In this paper, we combined gender and 13 well-confirmed, disease-associated loci to form a predictive genetic test for RA. The test reaches low to medium discriminative ability (AUC = 0.7). Although this may not reach the standard required for clinical purposes, it is still quite promising. The test has a higher discriminative ability than tests for other common complex diseases, such as type 2 diabetes (Lango et al., 2008; Lu et al., 2009b; Meigs et al., 2008; van Hoek et al., 2008) and prostate cancer (Zheng et al., 2008). With more common/rare variants, gene–gene/gene–environment interactions, and copy number variations to be discovered in the future, the test's accuracy could be further improved, and may eventually lead to a clinically useful test.

Acknowledgments

We thank two anonymous reviewers for their valuable comments and suggestions. This work was supported by the Intramural Research Grants Program (08-IRGP-1519-1) from Michigan State University, a U.S. Public Health Service Resource grant (RR03655) from the National Center for Research Resources, a research grant (GM28356) from the National Institute of General Medical Sciences, a Cancer Center Support Grant (P30CAD43703) from the National Cancer Institute, and a research grant (DMS 0707031) from the National Science Foundation. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for that project was provided by the Wellcome Trust, under award 076113.

REFERENCES

- Baker SG. Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics*. 2000; 56:1082–1087. [PubMed: 11129464]
- Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhardt AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van GA, Zelenika D, Franchimont D, Hugot JP, de VM, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* 2008; 40:955–962. [PubMed: 18587394]
- Bell J. The new genetics in clinical practice. *Br. Med. J.* 1998; 316:618–620. [PubMed: 9518921]
- Breiman L. Bagging predictors. *Machine Learning*. 1996; 24:123–140.
- Egan, JP. *Signal Detection Theory and ROC Analysis*. New York: Academic Press; 1975.
- Epstein CJ. Medical genetics in the genomic medicine of the 21st century. *Am. J. Hum. Genet.* 2006; 79:434–438. [PubMed: 16909381]

- Evans JP, Skrzynia C, Burke W. The complexities of predictive genetic testing. *Br. Med. J.* 2001; 322:1052–1056. [PubMed: 11325775]
- Gorman JD, David-Vaudey E, Pai M, Lum RF, Criswell LA. Particular HLA-DRB1 shared epitope genotypes are strongly associated with rheumatoid vasculitis. *Arthritis Rheum.* 2004; 50:3476–3484. [PubMed: 15529352]
- Guttmacher AE, Collins FS. Realizing the promise of genomics in biomedical research. *J. Am. Med. Assoc.* 2005; 294:1399–1402.
- Han TU, Bang SY, Kang C, Bae SC. TRAF1 polymorphisms associated with rheumatoid arthritis susceptibility in Asians and in Caucasians. *Arthritis Rheum.* 2009; 60:2577–2584. [PubMed: 19714643]
- Janssens AC, Pardo MC, Steyerberg EW, van Duijn CM. Revisiting the clinical validity of multiplex genetic testing in complex diseases. *Am. J. Hum. Genet.* 2004; 74:585–588. [PubMed: 14973786]
- Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L. The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genet. Med.* 2007; 9:665–674. [PubMed: 18073579]
- Lango H, Palmer CN, Morris AD, Zeggini E, Hattersley AT, McCarthy MI, Frayling TM, Weedon MN. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. *Diabetes.* 2008
- Lee AT, Li W, Liew A, Bombardier C, Weisman M, Massarotti EM, Kent J, Wolfe F, Begovich AB, Gregersen PK. The PTPN22 R620W polymorphism associates with RF positive rheumatoid arthritis in a dose-dependent manner but not with HLA-SE status. *Genes Immun.* 2005; 6:129–133. [PubMed: 15674368]
- Lu Q, Elston RC. Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. *Am. J. Hum. Genet.* 2008; 82:641–651. [PubMed: 18319073]
- Lu Q, Obuchowski N, Won S, Zhu X, Elston RC. Using the optimal robust receiver operating characteristic curve for predictive genetic tests. *Biometrics.* 2009a
- Lu Q, Song Y, Wang X, Won S, Cui Y, Elston RC. The effect of multiple genetic variants in predicting the risk of Type 2 Diabetes. *BMC Proc.* 2009b
- Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* 2005; 37:413–417. [PubMed: 15793588]
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 2007; 39:906–913. [PubMed: 17572673]
- McClish DK. Analyzing a portion of the ROC curve. *Med. Decision Making.* 1989; 9:190–195.
- McIntosh MW, Pepe MS. Combining several screening tests: Optimality of the risk score. *Biometrics.* 2002; 58:657–664. [PubMed: 12230001]
- Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB Sr, Cupples LA. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N. Engl. J. Med.* 2008; 359:2208–2219. [PubMed: 19020323]
- Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* 1996; 49:1373–1379. [PubMed: 8970487]
- Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Press; 2003.
- Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics.* 2006; 62:221–229. [PubMed: 16542249]
- Plenge RM, Padyukov L, Remmers EF, Purcell S, Lee AT, Karlson EW, Wolfe F, Kastner DL, Alfredsson L, Altshuler D, Gregersen PK, Klareskog L, Rioux JD. Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: Association of susceptibility with PTPN22, CTLA4, and PADI4. *Am. J. Hum. Genet.* 2005; 77:1044–1060. [PubMed: 16380915]

- Plenge RM, Cotsapas C, Davies L, Price AL, de Bakker PI, Maller J, Pe'er I, Burt NP, Blumenstiel B, Defelice M, Parkin M, Barry R, Winslow W, Healy C, Graham RR, Neale BM, Izmailova E, Roubenoff R, Parker AN, Glass R, Karlson EW, Maher N, Hafler DA, Lee DM, Seldin MF, Remmers EF, Lee AT, Padyukov L, Alfredsson L, Coby J, Weinblatt ME, Gabriel SB, Purcell SM, Klareskog L, Gregersen PK, Shadick NA, Daly MJ, Altshuler D. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.* 2007a; 39:1477–1482. [PubMed: 17982456]
- Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, Li W, Tan AK, Bonnard C, Ong RT, Thalamuthu A, Pettersson S, Liu C, Tian C, Chen WV, Carulli JP, Beckman EM, Altshuler D, Alfredsson L, Criswell LA, Amos CI, Seldin MF, Kastner DL, Klareskog L, Gregersen PK. TRAF1-C5 as a risk locus for rheumatoid arthritis-A genome-wide study. *N. Engl. J. Med.* 2007b; 357:1199–1209. [PubMed: 17804836]
- Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature.* 2009; 460:748–752. [PubMed: 19571811]
- Thomson W, Barton A, Ke X, Eyre S, Hinks A, Bowes J, Donn R, Symmons D, Hider S, Bruce IN, Wilson AG, Marinou I, Morgan A, Emery P, Carter A, Steer S, Hocking L, Reid DM, Wordsworth P, Harrison P, Strachan D, Worthington J. Rheumatoid arthritis association at 6q23. *Nat. Genet.* 2007; 39:1431–1433. [PubMed: 17982455]
- van Hoek M, Dehghan A, Witteman JC, van Duijn CM, Uitterlinden AG, Oostra BA, Hofman A, Sijbrands EJ, Janssens AC. Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. *Diabetes.* 2008; 57:3122–3128. [PubMed: 18694974]
- Vignat C, Bansal AT, Balding DJ, Binks MH, Dickson MC, Montgomery DS, Wilson AG. Genetic association of the major histocompatibility complex with rheumatoid arthritis implicates two non-DRB1 loci. *Arthritis Rheum.* 2009; 60:53–62. [PubMed: 19116923]
- Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007; 447:661–678. [PubMed: 17554300]
- Zervou MI, Sidiropoulos P, Petraki E, Vazgiourakis V, Krasoudaki E, Raptopoulou A, Kritikos H, Choustoulaki E, Boumpas DT, Goulielmos GN. Association of a TRAF1 and a STAT4 gene polymorphism with increased risk for rheumatoid arthritis in a genetically homogeneous population. *Hum. Immunol.* 2008; 69:567–571. [PubMed: 18625278]
- Zheng SL, Sun J, Wiklund F, Smith S, Stattin P, Li G, Adami HO, Hsu FC, Zhu Y, Balter K, Kader AK, Turner AR, Liu W, Bleecker ER, Meyers DA, Duggan D, Carpten JD, Chang BL, Isaacs WB, Xu J, Gronberg H. Cumulative association of five genetic variants with prostate cancer. *N. Engl. J. Med.* 2008; 358:910–919. [PubMed: 18199855]
- Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.* 1993; 39:561–577. [PubMed: 8472349]

Figure 1.

The four ROC curves in the plot, from top to bottom, were formed by the bagging optimal ROC curve method, logistic regression considering only single locus effects, the allele counting method, and logistic regression fitting single locus and interaction effects. The estimated AUC values from these four models are 0.7043, 0.7037, 0.6738, and 0.5564, respectively.

Table 1

Summary of 13 RA-associated loci

SNP	Gene	Chromosome	Source	Accuracy ^a
rs6457617	<i>MHC</i>	6	500K	NA
rs4678	<i>MHC</i>	6	imputed	0.9973
rs2442728	<i>MHC</i>	6	imputed	0.9849
rs660895	<i>HLA-DRB</i>	6	imputed	0.9970
rs6910071	<i>HLA-DRB</i>	6	imputed	0.9882
rs3817964	<i>HLA-DRB</i>	6	imputed	0.9853
rs2476601	<i>PTPN22</i>	1	imputed	0.9949
rs10818488	<i>TRAF1</i>	9	imputed	0.9999
rs6920220	<i>TNFAIP3</i>	6	500K	NA
rs2240340	<i>PADI4</i>	1	imputed	0.9999
rs3738919	<i>ITGAV</i>	2	imputed	0.9818
rs7574865	<i>STAT4</i>	2	imputed	0.9964
rs1061622	<i>TNFRSF1B</i>	1	imputed	0.9995

^aThe average maximum posterior genotype call probability for the given locus.

Table 2

A comparison of the AUC estimates and their standard deviations (SD) among the bagging optimal ROC curve method (BAG-OPT), logistic regression considering only single-locus effects (LOG-REG), the allele counting method (ALL-CNT), and logistic regression fitting both single-locus and interaction effects (INT LOG-REG) under a disease model without interaction

Method	<i>N</i> = 500		<i>N</i> = 1000	
	AUC	SD	AUC	SD
BAG-OPT	0.5854	0.0121	0.5876	0.0102
LOG-REG	0.5515	0.0358	0.5506	0.0294
ALL-CNT	0.5415	0.0228	0.5415	0.0171
INT LOG-REG	0.5385	0.0276	0.5321	0.0234

Table 3

A comparison among the bagging optimal ROC curve method (BAG-OPT), logistic regression considering only single-locus effects (LOG-REG), the allele counting method (ALL-CNT), and logistic regression fitting both single-locus and interaction effects (INT LOG-REG) under a disease model with gene–gene interaction

Method	<i>N</i> = 500		<i>N</i> = 1000	
	AUC	SD	AUC	SD
BAG-OPT	0.6153	0.0173	0.623	0.0148
LOG-REG	0.5876	0.0454	0.6005	0.0322
ALL-CNT	0.5782	0.0259	0.5800	0.0187
INT LOG-REG	0.5518	0.0351	0.5665	0.0326