Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.jzus.zju.edu.cn; www.springerlink.com
E-mail: jzus@zju.edu.cn

An ensemble-based likelihood ratio approach for family-based genomic risk prediction*#

Hui AN[1], Chang-shuai WEI[2], Oliver WANG[3], Da-hui WANG[1], Liang-wen XU[4], Qing LU[5], Cheng-yin YE[†‡1]

[1]Department of Health Management, School of Medicine, Hangzhou Normal University, Hangzhou 310036, China

[2]Department of Biostatistics and Epidemiology, University of North Texas Health Science Center, Fort Worth, TX 76107, USA

[3]HBI Solutions Inc, Palo Alto, CA 94301, USA

[4]Department of Preventive Medicine, School of Medicine, Hangzhou Normal University, Hangzhou 310036, China

[5]Department of Epidemiology and Biostatistics, College of Human Medicine, Michigan State University, East Lansing, MI 48824, USA

[†]E-mail: yechengyin@hznu.edu.cn

Received Mar. 14, 2018; Revision accepted July 12, 2018; Crosschecked Nov. 8, 2018

Abstract: Objective: As one of the most popular designs used in genetic research, family-based design has been well recognized for its advantages, such as robustness against population stratification and admixture. With vast amounts of genetic data collected from family-based studies, there is a great interest in studying the role of genetic markers from the aspect of risk prediction. This study aims to develop a new statistical approach for family-based risk prediction analysis with an improved prediction accuracy compared with existing methods based on family history. Methods: In this study, we propose an ensemble-based likelihood ratio (ELR) approach, Fam-ELR, for family-based genomic risk prediction. Fam-ELR incorporates a clustered receiver operating characteristic (ROC) curve method to consider correlations among family samples, and uses a computationally efficient tree-assembling procedure for variable selection and model building. Results: Through simulations, Fam-ELR shows its robustness in various underlying disease models and pedigree structures, and attains better performance than two existing family-based risk prediction methods. In a real-data application to a family-based genome-wide dataset of conduct disorder, Fam-ELR demonstrates its ability to integrate potential risk predictors and interactions into the model for improved accuracy, especially on a genome-wide level. Conclusions: By comparing existing approaches, such as genetic risk-score approach, Fam-ELR has the capacity of incorporating genetic variants with small or moderate marginal effects and their interactions into an improved risk prediction model. Therefore, it is a robust and useful approach for high-dimensional family-based risk prediction, especially on complex disease with unknown or less known disease etiology.

Key words: Family-based study; Genetic risk prediction; High-dimensional data
https://doi.org/10.1631/jzus.B1800162 **CLC number:** Q39

1 Introduction

With rapidly evolving high-throughput technologies, very large numbers of genetic markers have been genotyped for the discovery of new disease-associated variants in family studies. While the novel findings from family-based association studies likely further improve our understanding of disease etiologies, the genetic data collected for these studies also provide us with a great opportunity to systematically study the role of vast numbers of genetic markers in

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 81402762) and the National Institute on Drug Abuse (Nos. K01DA033346 and R01DA043501), USA

Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/jzus.B1800162) contains supplementary materials, which are available to authorized users

ORCID: Cheng-yin YE, https://orcid.org/0000-0001-8384-1857

family-based risk prediction. The hope is that ultimately we can incorporate the genetic information into clinical practice for the early identification of disease susceptibility and individualized preventive strategies (Ginsburg and Willard, 2009; Abraham and Inouye, 2015).

Studies have been initiated for genetic risk prediction of human diseases, such as type 2 diabetes, cardiovascular diseases, cancers, and psychiatric disorders (e.g. schizophrenia, autism, and bipolar disorder) (Wray et al., 2014; Smith et al., 2015; Choi et al., 2016; Shieh et al., 2016). Being different from most Mendelian disorders, where causal genetic variants have almost complete penetrance, complex diseases are likely caused by complex interplay of genetic and environmental risk factors, and thus their genetic etiology is largely unknown. Since the number of uncovered genetic and environmental risk factors is still limited, it remains a great challenge to form an accurate risk prediction model for most complex diseases. As a result, most existing risk prediction models that are built on a handful of known risk predictors have low accuracy for potential clinical use. These models can be potentially improved by additional risk predictors, such as variants with small- or medium-effect sizes, as evidence has shown that collectively small- or medium-effect variants can explain a large proportion of disease variations (Yang et al., 2010).

The genetic risk score (GRS) approach is one of the most widely used and easily implemented methods for genetic risk prediction. It usually assumes an additive genetic disease model and forms an overall genotypic risk score for prediction by summing risk alleles across multiple disease-associated loci, either with or without weighting on the effect sizes of loci. Although this weighting strategy could relax the method's assumption of equal effect sizes among effective genetic variants, the empirical calculation could also cause bias when potential variations occur across different studies. Furthermore, with the assumption that variants involved have to be effective and independent of each other, the GRS method could be subject to low performance if non-causal variants or interactions exist (Chatterjee et al., 2013). An extension of the GRS approach, called GRS-based generalized estimating equation (GS-GEE), has been developed for family-based data (Meigs et al., 2008). As with other GRS approaches, GS-GEE has the limitations of not considering interactions and being less robust to noise signals. Few approaches have been developed for family-based risk prediction analysis. Besides GS-GEE, Bayesian Lasso approaches (de los Campos et al., 2009), widely used in animal breeding, can also be used for family-based risk prediction. More recently, a random field method has been developed for family-based risk prediction (Wen et al., 2017). Although some of those methods, such as Lasso and random field, could be applied to genome-wide data technically, they are still commonly adopted to build gene-based risk models, where interactions within each gene unit may be considered, but interactions across different gene units may still fail to be detected.

In this paper, we propose a nonparametric approach—an ensemble-based likelihood ratio (ELR) approach—for family-based risk prediction research, Fam-ELR. This study extends our previously developed family-based risk prediction method, clustered optimal receiver operating characteristic (ROC) curve (CORC) (Ye et al., 2011a). CORC incorporated a clustered ROC curve method to consider sample correlations on a family-based dataset, and adopted a computationally efficient forward-searching algorithm for risk model construction. In addition to features inherited from CORC, Fam-ELR uses a tree-assembling process to simultaneously assemble numerous risk prediction trees. This new approach can potentially attain better performance than the previous approach by considering a large number of genetic variants with small/moderate marginal effects and their possible within and between gene interactions. Simulation studies were conducted to compare Fam-ELR's performance with GS-GEE and CORC, and to assess the method's robustness among various disease models and pedigree structures. Finally, we applied Fam-ELR to the family-based genome-wide data of conduct disorder (CD), studying two CD risk prediction models, one based on known CD risk predictors and the other based on genome-wide genetic markers.

2 Methods

Assume in a family-based dataset G^p_{ij}=(g_{ij1}, g_{ij2}, …, g_{ijp}) is the p-dimensional risk profile for the jth individual (j=1, 2, …, m_i, $\sum_{i=1}^{N} m_i = M$) from the

ith family (i=1, 2, …, N), which belongs to one of p_K p-dimensional risk profiles ($G^p_{ij} \in G^p_k$, k=1, 2, …, p_K). y_{ij} denotes the binary measurement of an interested phenotype, such as a disease status (e.g. $y_{ij} \in S$, S=1 for a disease status and S=0 for a non-disease status). In our previously proposed CORC approach, we first calculate the probabilities of p_K risk profiles conditional on disease status, $P(G^p_{ij}|S=1)$ and $P(G^p_{ij}|S=0)$, and then derive the likelihood ratio (LR) using the equation $\mathrm{LR}(G^p_{ij})=P(G^p_{ij}|S=1)/P(G^p_{ij}|S=0)$. Based on the LRs, we obtain the clustered area under the curve (AUC) value of the model, $\mathrm{AUC}^p_{\mathrm{cluster}}$, using the algorithm proposed by Obuchowski (1997) to take within-family correlations into consideration. Generally speaking, the Obuchowski's method is a nonparametric algorithm and does not make any assumptions about the intra-family correlation structure. It first gives equal weight to all pairwise rankings within and between families, and then separates the scores of individuals into two distinct components (i.e. affected- and unaffected-components). By doing this, the algorithm derives the variance of the clustered AUC by taking into consideration not only the variance of both affected- and unaffected-components across families, but also the correlation between the two components within each family.

With the goal of integrating hundreds of potential risk predictors, as well as their interactions, into an improved risk prediction model, a tree-assembling process is used. Suppose p genetic variants were genotyped for the M individuals in N families. By treating each family as a sampling unit, we draw T (e.g. 1000) bootstrap populations from the original population, each bootstrap population consisting of N families, and obtain T corresponding out-of-bag populations. For each bootstrap sample, a forward selection algorithm is implemented to build a tree-based risk prediction model (Ye et al., 2011b). The forward selection algorithm starts with a null model of no predictors. It gradually selects potential risk predictors into the model by searching exhaustively among available genetic variants, and keeps splitting the samples into different risk groups in a binary fashion. In each step, the variant that adds the highest accuracy to the model is selected into the model. The algorithm continues until a prediction model is complete. By applying the forward selection algorithm to all T bootstrap samples, we construct a large ensemble of risk prediction models, each containing a collection of diverse but potentially useful risk predictors, some with low- or medium-marginal effects. By applying this ensemble of tree-based models to the corresponding out-of-bag samples, we can calculate the LR values. The LR value for the jth individual from the ith family can be obtained by averaging its LR values across all out-of-bag samples. The averaged LR value, $\overset{\frown}{\mathrm{LR}}_{ij}$, is used as the risk score (i.e. $\overset{\frown}{\mathrm{LR}}{}^{\mathrm{a}}_{ij}$ for affected individuals and $\overset{\frown}{\mathrm{LR}}{}^{\mathrm{u}}_{ij}$ for unaffected individuals) to calculate the averaged clustered AUC value, $\mathrm{AUC}^{\mathrm{A}}_{\mathrm{cluster}}$, by

$$\mathrm{AUC}^{\mathrm{A}}_{\mathrm{cluster}}=\frac{1}{AU}\sum_{i=1}^{N}\sum_{i'=1}^{N}\sum_{l=1}^{a_i}\sum_{k=1}^{u_{i'}}\varphi(\overset{\frown}{\mathrm{LR}}{}^{\mathrm{a}}_{il},\overset{\frown}{\mathrm{LR}}{}^{\mathrm{u}}_{i'k}),$$

where

$$A=\sum_{i}^{N}a_i,\ \ U=\sum_{i}^{N}u_i,$$

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

The variance of the $\mathrm{AUC}^{\mathrm{A}}_{\mathrm{cluster}}$ can also be obtained by

$$\mathrm{var}(\mathrm{AUC}^{\mathrm{A}}_{\mathrm{cluster}})=\frac{N_a\sum_{i=1}^{N_a}[V_a(\overset{\frown}{\mathrm{LR}}{}^{\mathrm{a}}_{i.})-a_i\mathrm{AUC}^{\mathrm{A}}_{\mathrm{cluster}}]^2}{A^2(N_a-1)}+\frac{N_u\sum_{i=1}^{N_u}[V_u(\overset{\frown}{\mathrm{LR}}{}^{\mathrm{u}}_{i.})-u_i\mathrm{AUC}^{\mathrm{A}}_{\mathrm{cluster}}]^2}{U^2(N_u-1)}+\frac{2N\sum_{i=1}^{N}[V_a(\overset{\frown}{\mathrm{LR}}{}^{\mathrm{a}}_{i.})-a_i\mathrm{AUC}^{\mathrm{A}}_{\mathrm{cluster}}][V_u(\overset{\frown}{\mathrm{LR}}{}^{\mathrm{u}}_{i.})-u_i\mathrm{AUC}^{\mathrm{A}}_{\mathrm{cluster}}]}{AU(N-1)},$$

where

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

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

By aggregating subsets of different genetic risk predictors, the new method can simultaneously consider a large number of potential risk predictors, especially those with relatively small marginal effects, and their possible interactions. This could further improve the model's accuracy.

3 Results

3.1 Simulation studies

We conducted two simulations to evaluate the performance of Fam-ELR by comparing its prediction accuracy and robustness with those of two existing family-based risk prediction approaches, GS-GEE and CORC. The commonly used GS-GEE approach first calculates a summarized genotype risk score by counting the number of risk alleles and then adopts a generalized estimating equation method to build a risk prediction model with consideration of the familiar correlation. CORC is an approach we previously developed for family-based risk prediction analysis. It integrates a clustered ROC curve method into a computationally efficient forward algorithm. In the first simulation, we simulated different disease models by varying the total number of disease-associated variants, modes of inheritance, and types of interactions. In the second simulation, we varied pedigree structures from simple trios to complicated three-generation pedigrees, and evaluated the impact of family structures on the methods' performance. For each simulation setting, we generated 1000 replicates, and split all samples into a training set and a validation set with a 2:1 ratio. The training set was used to construct the risk prediction model, while the validation set served as independent data to evaluate the performance of the model and estimate the AUC.

3.1.1 Simulation under various disease models

The ideas of this simulation study are based on the hypothesis that our proposed Fam-ELR method can potentially attain better performance than previous approaches (i.e. CORC and GS-GEE) when a large number of genetic variants with small to moderate marginal effects and possible interactions within and across genes are involved in disease models. Therefore, in order to demonstrate and evaluate our method's prediction ability in such situations, we gradually modified the model's complexity not only by gradually increasing the number of disease-associated single nucleotide polymorphisms (SNPs) and their interactions, but also by changing the underlying interaction types from two-way interactions to three-way interactions and changing from the threshold mode to the multiplicative mode. The population disease prevalence ranges from 0.045 to 0.049. Under each disease model, 10 noise SNPs were included, with the allele frequencies ranging from 0.2 to 0.8. Specifically, model 1 included 10 noise SNPs and 5 additive-effect SNPs with the odds ratios ranging from 1.5 to 1.9. In model 1, we also simulated a two-way threshold interaction between 2 of 5 additive-effect SNPs. While maintaining the number of noise SNPs at 10 in model 2, we increased the number of disease-associated SNPs to 10 and generated 5 two-way threshold interaction models across 10 disease-associated SNPs. Model 3 is similar to model 2 except that the underlying interaction model is the multiplicative model instead of the threshold interaction model. In model 4, we increased the number of disease-associated SNPs to 12 and simulated 2 two-way threshold interactions and 2 three-way threshold interactions. The numbers of SNPs and interactions, as well as details of threshold and multiplicative interactions, were designed based on the multi-locus theories of common complex diseases illustrated in previous studies (Marchini et al., 2005; Wei et al., 2013).

For this simulation scenario, we first generated a population of 1000000 samples, and then simulated nuclear families with two parents and two offspring. Based on individuals' genotypes and disease models, we simulated the phenotypes. A total of 1000 replicates were simulated for each disease model. For each replicate, 1000 nuclear families with at least one affected family member were sampled from the population. The true AUC, also defined as the expected AUC, is the measure of the expected discriminative ability of a prediction model/test and is uniquely determined from each unique disease model by its simulation settings. These settings of the underlying genetics, including the number and the effect size of causal SNPs, their genotype distributions and interaction modes, will generate true risk scores for the population. These are then used to compute the expected AUC. To mimic common disease scenarios, we set the simulation parameters such that the expected AUCs in our simulations have moderate values of around 0.76 (listed in Tables 1 and 2), based on the hypothesis that most common diseases are prone to being caused by the simultaneous interplay of hundreds of genetic and environmental risk factors, and the genetic component plays a relatively minor or moderate marginal or interactive role, and thus the expected AUC of a genetic predictive test is usually lower than 0.8 (Janssens and van Duijn, 2008).

The results of simulation are summarized in Table 1 and Fig. 1, where predicted-AUC means (mean), standard deviations (SDs), bias from the true AUC values (bias), and mean square errors (MSEs) are reported. After running 1000 simulations for a certain simulation setting, the bias is calculated as the distance between the mean of estimated AUC and the true AUC (i.e. bias=mean of estimated AUC−true AUC) (Wackerly et al., 2008). When the bias achieves a negative value, it is an indication that the target algorithm underestimates the true AUC and does not suffer from overfitting. Therefore, in this simulation, all three algorithms do not show signs of overfitting. In addition, our proposed Fam-ELR method always attains the biases closest to zero, and thus it has a better estimate of the true AUC than CORC and GS-GEE. For instance, with 10 low-effect SNPs and 5 two-way threshold interactions, model 2 tends to mimic a common disease situation where risk predictors are subjected to low- to medium-effects with possible interactions. The results of model 2 reveal that our proposed Fam-ELR method has an AUC mean of 0.6723, a 3.85% and 12.90% increase in accuracy on that of the CORC method (AUC mean= 0.6474) and the GS-GEE method (AUC mean= 0.5955), respectively. Moreover, Fam-ELR also attains smaller MSEs than the other two algorithms in these simulations, implying that the proposed method achieved a better overall performance than the other two algorithms when taking into account both prediction errors (the bias) and deviations (the variance). MSE is a commonly used summary estimator of prediction quality and always has a non-negative value. The smaller the value is, the better the achieved performance is (Wackerly et al., 2008). In model 2, the MSE of Fam-ELR is 0.0078, 39.53% and 71.11% lower than the MSEs of CORC and GS-GEE, respectively, indicating higher accuracy and lower variation of Fam-ELR. As shown in Fig. 1, when the disease model is simple, with limited SNPs and interactions (i.e. model 1), Fam-ELR and CORC perform similarly, with only a slight difference between their MSEs (i.e. MSE=0.0061 for Fam-ELR and MSE=0.0071 for CORC). However, when the model complexity increases along with the interaction settings (i.e. models 2, 3, and 4), Fam-ELR remains robust and accurate (its MSE consistently around

Table 1 Performance of Fam-ELR, CORC, and GS-GEE under four disease models*

Method	Mean	Bias	SD	MSE
Model 1 (true AUC=0.7567)				
Fam-ELR	0.6799	−0.0768	0.0148	0.0061
CORC	0.6740	−0.0827	0.0158	0.0071
GS-GEE	0.5854	−0.1713	0.0139	0.0295
Model 2 (true AUC=0.7594)				
Fam-ELR	0.6723	−0.0871	0.0153	0.0078
CORC	0.6474	−0.1120	0.0188	0.0129
GS-GEE	0.5955	−0.1639	0.0139	0.0270
Model 3 (true AUC=0.7576)				
Fam-ELR	0.6451	−0.1125	0.0156	0.0129
CORC	0.5974	−0.1603	0.0194	0.0261
GS-GEE	0.6237	−0.1339	0.0138	0.0181
Model 4 (true AUC=0.7879)				
Fam-ELR	0.6885	−0.0994	0.0127	0.0100
CORC	0.6723	−0.1156	0.0155	0.0136
GS-GEE	0.5993	−0.1886	0.0142	0.0358

* Model 1 involves 5 SNPs with 1 two-way threshold interaction; model 2 involves 10 SNPs with 5 two-way threshold interactions; model 3 involves 10 SNPs with 5 two-way multiplicative interactions; and model 4 involves 12 SNPs with 2 two-way threshold interactions and 2 three-way threshold interactions

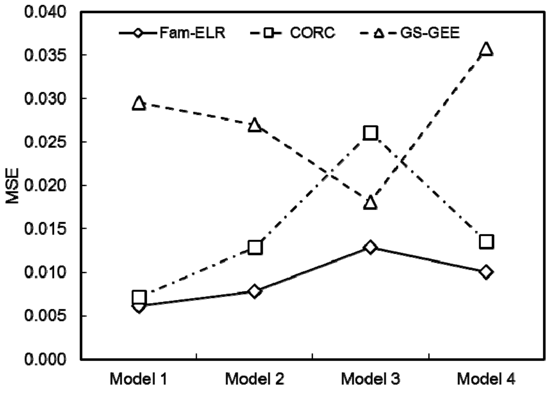

Fig. 1 Mean square errors (MSEs) of Fam-ELR, CORC, and GS-GEE under four disease models

MSE=bias^2+variance

0.01), whereas the performance of the other two methods varies across different models (i.e. MSEs of CORC and GS-GEE vary from 0.0129 to 0.0261 and 0.0181 to 0.0358, respectively), and is all inferior to that of Fam-ELR. In summary, our simulations show that the proposed Fam-ELR algorithm can utilize a tree-assembling process to integrate variants with small to medium effects, as well as their interactions, and ultimately attain better performance than that of previous approaches.

3.1.2 Simulation under various pedigree structures

To investigate the impact of pedigree structure on the performance of family-based risk prediction methods, three different pedigrees were evaluated in this simulation: trios with two parents and an affected child, four-member nuclear families with at least one affected family member, and three-generation pedigrees with a total of 10 family members. Two disease models were considered for each pedigree setting: the two-way threshold model (model 2) and the two-way multiplicative model (model 3) used in simulation 1. In this simulation, 1200 trios, 1000 nuclear families, and 450 three-generation pedigrees were generated from the 1000000-sample population, with sample sizes ranging from 3600 to 4500.

The simulation results are summarized in Table 2. Overall, the proposed Fam-ELR method attains the highest accuracy among the three methods. For example, under the two-way multiplicative model and the three-generation pedigree, the classification accuracy of Fam-ELR (AUC=0.6686) is higher than those of the CORC (AUC=0.6084) and the GS-GEE (AUC=0.6532) methods. Fam-ELR also obtains an MSE of 0.0086, which is 62.9% and 25.2% lower than those of CORC (MSE=0.0232) and GS-GEE (MSE=0.0115), respectively. Fig. 2 shows Fam-ELR consistently has the lowest MSE in all simulation settings, indicating a more robust performance of Fam-ELR than that of the other two methods. Another interesting phenomenon observed from Fig. 2 is that CORC performs better than GS-GEE under the two-way threshold interactive model while GS-GEE outperforms CORC under the two-way multiplicative interactive model. The underperformance of GS-GEE under the two-way threshold interactive model can be explained by the violation of the additive assumption in GS-GEE.

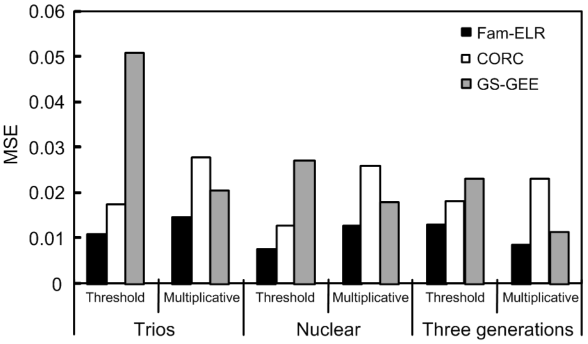

Fig. 2 Barplot of mean square errors (MSEs) of Fam-ELR, CORC, and GS-GEE under three pedigree settings and two disease models

Table 2 Performance of Fam-ELR, CORC, and GS-GEE under different family structures and disease models*

Method	Trios				Nuclear families				Three-generation pedigree			
	Mean	Bias	SD	MSE	Mean	Bias	SD	MSE	Mean	Bias	SD	MSE
Model 2 (true AUC=0.7594)												
Fam-ELR	0.6556	−0.1038	0.0143	0.0110	0.6723	−0.0871	0.0153	0.0078	0.6482	−0.1112	0.0267	0.0131
CORC	0.6285	−0.1309	0.0196	0.0175	0.6474	−0.1120	0.0188	0.0129	0.6279	−0.1315	0.0297	0.0182
GS-GEE	0.5420	−0.2174	0.0599	0.0508	0.5955	−0.1639	0.0139	0.0270	0.6087	−0.1507	0.0223	0.0232
Model 3 (true AUC=0.7576)												
Fam-ELR	0.6372	−0.1204	0.0137	0.0147	0.6451	−0.1125	0.0156	0.0129	0.6686	−0.0891	0.0251	0.0086
CORC	0.5917	−0.1659	0.0182	0.0279	0.5974	−0.1603	0.0194	0.0261	0.6084	−0.1492	0.0303	0.0232
GS-GEE	0.6146	−0.1430	0.0113	0.0206	0.6237	−0.1339	0.0138	0.0181	0.6532	−0.1044	0.0246	0.0115

* Model 2 involves 10 SNPs with 5 two-way threshold interactions and model 3 involves 10 SNPs with 5 two-way multiplicative interactions

3.2 Family-based risk prediction analysis of conduct disorder

Most risk prediction studies construct risk prediction models by focusing on previously reported disease-susceptibility genetic and environmental risk factors. Such procedures can achieve remarkable success under certain disease scenarios, especially when diseases are caused by a limited number of genetic and environmental risk factors, each associated with a large effect. For instance, a risk prediction model of age-related macular degeneration (AMD) builds on five major AMD-associated variants and can reach a high accuracy (AUC=0.8) (Maller et al., 2006). Nevertheless, most common diseases are likely caused by interplay of hundreds of genetic and environmental risk factors, each with a low or medium marginal effect. The performance of risk prediction models for such diseases can be improved by including not only known risk factors but also other potential risk predictors. We adopted both strategies in our real data application. Comparing the simulations that mainly focused on our algorithm's performance on various disease settings with distinct interaction modes, we further demonstrated our algorithm's ability for high-dimensional disease risk prediction in the real data analysis. We started this with a risk prediction analysis of known predictors of CD, and then extended the prediction to the genome-wide scale, searching for new risk predictors to further improve the model's accuracy. In both analyses, we evaluated the performance of three methods.

CD is a serious behavioral and emotional disorder of children and teens. A child affected by CD may display a set of disruptive and violent behaviors, and thus fail to obey rules and further violate the rights of others, having a severe influence on the family's and child's own daily life (Kazdin, 1997). For a disease that is prevalent among children and teens, it is quite popular to use a family-based study design strategy to investigate genetic risk factors and family environment related to CD. For this analysis, we use samples from the International Multicenter ADHD (attention-deficit/hyperactivity disorder) Genetics Project (the IMAGE project). The main purpose of the IMAGE project is to investigate the genetic causes of both CD and ADHD. The IMAGE dataset includes 206 CD cases and 2520 controls from over 900 parent-child trios.

3.2.1 Family-based risk prediction analysis of known CD predictors

For the real data application, we first performed a family-based risk prediction analysis based on gender and 46 previously reported CD-associated loci (Ye et al., 2011a), and by using Fam-ELR, CORC, and GS-GEE. We treated each family as a unit, and split the whole dataset into a training dataset and a validation dataset with a ratio of 2:1. The training data were used for model building, while the validation dataset was used for model evaluation. The ROC curves of the models from the training and the validation datasets are plotted in Fig. 3. Based on the results

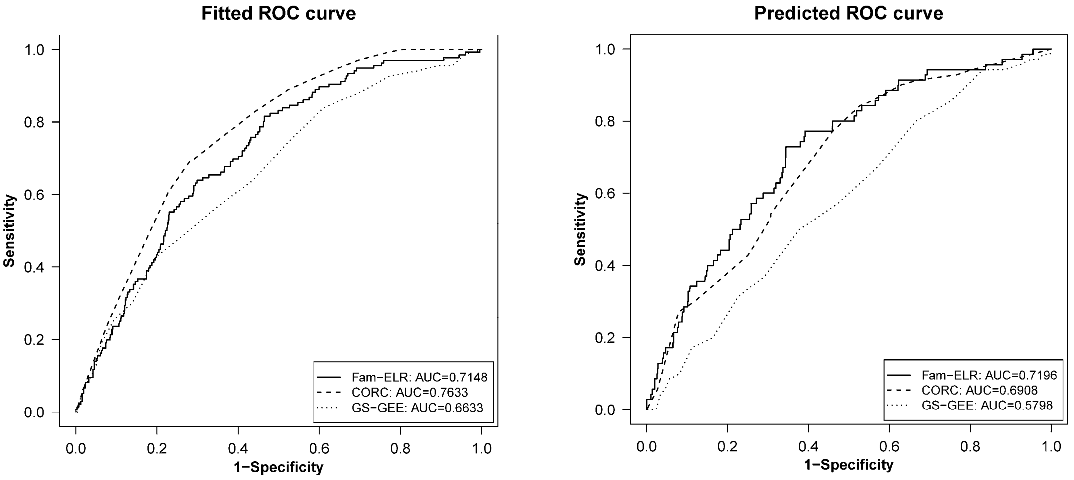

Fig. 3 ROC curves of risk prediction models formed by Fam-ELR, CORC, and GS-GEE based on known CD risk predictors

from the validation dataset, Fam-ELR achieved the highest prediction accuracy (i.e. AUC=0.7196), while the GS-GEE method had the lowest accuracy (AUC= 0.5798). The CORC method with an AUC of 0.6908 has a slightly lower accuracy than Fam-ELR. The small difference in accuracy between Fam-ELR and CORC could be explained by the shared forward selection algorithm and a relatively small set of CD-known predictors. This result is consistent with our simulation study, which also shows the similar performance of Fam-ELR and CORC under simple disease scenarios involving a limited number of risk predictors.

We further explore the results from Fam-ELR, and summarize the number of times a predictor selected by the forward selection algorithm. In total, our proposed algorithm recruited 17 SNPs as predictors with non-zero weights. The top 10 most important predictors are summarized in Table 3. The top predictor is gender, and the remaining top-ranked predictors are *rs10831284*, *rs10492664*, *rs10229603*, *rs1644305*, *rs10797919*, *rs2826340*, *rs7595103*, *rs6427356*, and *rs2825388*, all of which have been reported as being significantly associated with CD in previous genome-wide association studies (GWAS) (Anney et al., 2008; Sonuga-Barke et al., 2008). Compared with Fam-ELR, only 4 predictors, gender, *rs10492664*, *rs10797919*, and *rs1644305*, are selected by CORC. These predictors are also among the top predictors in Fam-ELR, and rank as the 1st, 3rd, 6th, and 5th among all predictors in the Fam-ELR model. Since GS-GEE assumes equal effects of all predictors, we are not able to study the relative importance of predictors in the GS-GEE model. In the following exploratory analysis, we adopted logistic regression to further investigate the relationship among those impactful predictors, wherein a significant two-way interaction between *rs10229603* and *rs1644305* was found (with a *P*-value of 0.0340), implying that Fam-ELR, as a tree-based method, is able to take potential interaction effects into account.

3.2.2 Genome-wide family-based risk prediction analysis

We further conducted a genome-wide family-based risk prediction analysis to explore additional predictors that can be used to improve CD risk prediction. For this analysis, we only compared Fam-ELR and CORC because both approaches were designed for high-dimensional risk prediction analysis involving a large number of predictors and excluding noise signals. Before the risk prediction analysis, we performed a quality control analysis. After removing SNPs with low calling rate, low minor allele frequency (MAF), and the departure from Hardy-Weinberg Equilibrium (*P*-value of $<1\times10^{-6}$), a total of 288925 SNPs remained in the genome-wide analysis. For the genome-wide risk prediction analysis, we split the samples into training and validation datasets. The models were formed in the training dataset and were then evaluated in the validation dataset. Due to the large number of SNPs, we also adopted a simple filtering procedure (Wei et al., 2012) to remove a large number of noise predictors. For the filtering process, we used the transmission disequilibrium test to perform a univariate screening. By varying the *P*-value cutoffs, we selected a subset of SNPs and filtered out those SNPs with a *P*-value larger than the *P*-value cutoff. By applying Fam-ELR and CORC to the

Table 3 Top 10 risk predictors selected by Fam-ELR from the family-based risk prediction analysis of known CD predictors

Rank	Predictor	Chromosome	Function	Gene	Rank in CORC
1	Gender				1
2	*rs10831284*	11	Regulatory region	*AMOTL1*, *CWC15*, *JMJD2D*	
3	*rs10492664*	13	Intergenic	*LIG4*, *ABHD13*	2
4	*rs10229603*	7	Intron	*FLJ31818*, *GPR85*	
5	*rs1644305*	5	Intergenic	*c5orf15*	4
6	*rs10797919*	1	Splice region of *RGL1*	*GLT25D2*	3
7	*rs2826340*	21	Intergenic		
8	*rs7595103*	2	Intergenic		
9	*rs6427356*	1	Intergenic	*ETV3L*, *ETV3*	
10	*rs2825388*	21	Intergenic	*PPIAP22*, *SLC6A6P1*	

different subset of SNPs, we formed risk-prediction models, and then evaluated these risk-prediction models on the validation datasets. Among those models, the model with *P*-value cutoff of 1×10^{-3} attains the highest accuracy. In this model, a total of 251 SNPs passed the *P*-value threshold and were used to build the CD risk prediction models. The ROC curves of the models in the training and validation datasets are presented in Fig. 4. The results from the validation dataset indicate that, based on more risk predictors, the model formed by Fam-ELR (AUC=0.8829) attains a much higher accuracy than that by CORC (AUC=0.6871). Compared to the analysis on a limited number of known CD predictors, the genome-wide analysis results in more accurate risk prediction models. Moreover, the results also indicate that Fam-ELR has an advantage over the other methods in the high-dimensional risk prediction analysis involving a large number of predictors with small or medium effects.

We further studied two risk prediction models formed by CORC and Fam-ELR. The model formed by CORC has three risk predictors: gender, *rs4546404*, and *rs184817*, which are also the top three predictors selected by Fam-ELR. Besides these three risk predictors, Fam-ELR also captured risk predictors with small or medium effects. The resulting algorithm captured 146 SNPs as predictors with non-zero effects. In Table 4, we summarize the top 20 risk predictors selected by Fam-ELR. Among them, some of the predictors were previously found to be associated with CD, such as gender, *rs10492664* between *LIG4* and *ABHD13*, and *rs7595103* within *LOC101927967* (Anney et al., 2008; Sonuga-Barke et al., 2008). *LOC101927967* has also been found to be associated

Table 4 Top 20 risk predictors selected by Fam-ELR from the genome-wide risk prediction analysis

Rank	Predictor	Chr	Gene	Rank in CORC
1	Gender			1
2	*rs4546404*	5	*LOC105377700*	2
3	*rs184817*	2	*NPAS2*	3
4	*rs10492664*	13	*LIG4*, *ABHD13*	
5	*rs12910488*	15		
6	*rs1389660*	3	*ZNF385D*	
7	*rs13058781*	3	*SLC6A6*	
8	*rs8002852*	13		
9	*rs1487044*	2	*LOC101927967*	
10	*rs11647668*	16	*LOC105371393*	
11	*rs2708919*	2		
12	*rs870488*	5		
13	*rs755101*	6	*SNAP91*	
14	*rs1882668*	12		
15	*rs1317508*	12		
16	*rs4078017*	1	*PFDN2*	
17	*rs17636733*	15	*UBE3A*, *LOC105370737*, *ATP10A*	
18	*rs7595103*	2	*LOC101927967*	
19	*rs2833834*	21	*EVA1C*	
20	*rs9396888*	6	*LOC105374956*	

Chr: chromosome

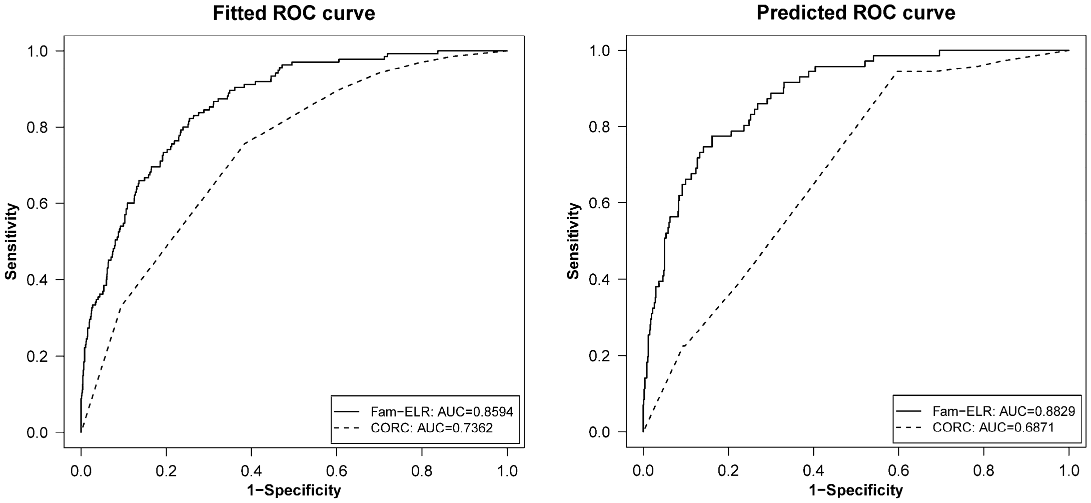

Fig. 4 ROC curves of risk prediction models formed by Fam-ELR and CORC based on CD genome-wide data

with cannabis dependence, childhood and early adolescence aggressive behavior in other studies (Pappa et al., 2016; Sherva et al., 2016). This may suggest a genetic overlap between CD and cannabis dependence. Some of the top 20 predictors have never been reported as directly associated with CD. They are, however, located within genes having strong impact on diseases related to CD. For instance, the 3rd and 7th selected loci, *rs184817* and *rs13058781*, are located within genes *NPAS2* and *SLC6A6*, respectively. *NPAS2* and *SLC6A6* have been reported as associated with cognitive performance, which might indicate their potential role in CD (Need et al., 2009; Rietveld et al., 2014). The selected predictors *rs1389660*, *rs1487044*, and *rs755101* are located in genes *ZNF385D*, *LOC101927967*, and *SNAP91*, respectively. Among them, *ZNF385D* is associated with ADHD and bipolar disorders (Ferreira et al., 2008; Lasky-Su et al., 2008), *LOC101927967* is known as an important ADHD- and cannabis-dependence-related gene (Anney et al., 2008; Sherva et al., 2016) which is also associated with childhood and early adolescence aggressive behavior (Pappa et al., 2016), while *SNAP91* has been reported to play a role in the morbidity of bipolar disorder with mood-incongruent psychosis and schizophrenia (Goes et al., 2012, 2015). While further studies are required to confirm the role of these selected predictors in CD, there is evidence indicating that Fam-ELR is able to capture known CD-associated variants as well as new variants potentially related to CD. By integrating new risk predictors and known CD predictors into the model, Fam-ELR improves the model's performance. By further exploring the top 20 risk predictors using logistic regression, we identified 31 two-way interactions as being significant (Table S1), showing the signs of latent interaction effects in the genome-wide prediction model constructed by Fam-ELR.

4 Discussion

The use of human genome discoveries and other established risk predictors for early disease prediction is an essential step towards precision medicine. However, the task of developing clinically useful risk prediction models is hampered by the present state of evidence, in which currently known risk predictors are insufficient for accurately predicting most human diseases. It has been shown that integrating predictors with small to medium effects into the risk prediction model could substantially improve a model's accuracy. In this study, we propose Fam-ELR for family-based risk prediction analysis. This proposed approach shares several unique features with our previous developed approaches (e.g. being applicable to various pedigree structures). In addition, it utilizes a tree-assembling process to integrate variants with small to medium effects, as well as their possible interactions, into the model for improved accuracy. Therefore, it offers a useful tool for high-dimensional risk prediction (e.g. genome-wide risk prediction).

Complex diseases are likely influenced by interplay of genetic and environmental risk predictors with an unknown underlying disease mechanism. While non-parametric methods are computationally efficient and rely on fewer assumptions about disease models, they have been less developed for high-dimensional risk prediction analysis, especially for family-based studies. Fam-ELR is a non-parametric approach that makes no assumption on underlying disease models and adopts a computationally efficient forward-searching algorithm for high-dimensional genetic data analysis. The results of the simulations reveal that Fam-ELR achieves a more robust and accurate performance than the other two methods, with lower MSEs and higher AUCs, regardless of the disease models and pedigree structures. Therefore, Fam-ELR is shown to be a robust approach for family-based risk prediction on complex disease with unknown or less known disease etiology. The simulation result also indicates that the commonly used GS-GEE attains good performance when the disease model follows a multiplicative model, but low performance when the underlying disease model is not multiplicative (e.g. a threshold model). The varied performance of GS-GEE under different models could be due to its additive/multiplicative assumption. It is also worth noting that, the biases in simulations all achieve negative values, revealing that the predicted values are lower than the true values. It indicates that the three algorithms are immune to the overfitting issue but tend to underestimate the true AUC.

On the other hand, compared to the commonly used random forest algorithm, our Fam-ELR method adopts the forward selection strategy to search for impactful predictors, whereas random forest usually selects significant features from a randomly selected subset of variants. Both strategies are feasible on the high-dimensional risk prediction and have their unique characteristics. Random forest would be more computationally efficient and tend to select variants with a relatively small effect size. However, random forest may be subject to low accuracy when a large proportion of loci are noise loci, which might be expected in the genome-wide risk prediction scenario (Ye et al., 2011b), while our Fam-ELR method is less sensitive to a large number of noise loci. In our study, although a variant-filtering process was introduced before the real-data genome-wide risk prediction, it is not a prerequisite step for our method. Furthermore, our method is designed to construct risk prediction based on a family-based dataset, while random forest is a commonly used case-control-based method.

In this study, Fam-ELR was applied to a family-based GWAS dataset on CD, utilizing both a limited number of known risk variants and the genome-wide level data to construct risk prediction models. For the risk prediction analysis on a limited number of known risk predictors, the performance of Fam-ELR and CORC is quite similar, indicating a limited advantage of Fam-ELR over other methods. Nevertheless, for the genome-wide risk prediction analysis, Fam-ELR has a significant advantage over CORC as it considers low- and medium-effect predictors via the tree-assembling procedure. In the genome-wide risk prediction analysis, many of these low- and medium-effect predictors have never been reported before, but are located within genes or genomic regions associated with CD or its related mental and behavioral disorders, such as cognitive performance, ADHD, and bipolar disorder. Such variants might have failed to pass the stringent significance threshold in the GWAS studies, but can have a significant predictive value in predicting CD. By considering these predictors, Fam-ELR could improve the model's accuracy performance on the high-dimensional scale.

The usefulness of a risk prediction model varies by the disease's prevalence, the availability of prevention and intervention methods, and the cost of surveillance measures. Furthermore, risk prediction based on genomic information should be treated as a complementary and integrated part of a more accurate prediction procedure that also considers other omic data, clinical biomarkers, and environmental risk factors. As the high-throughput sequencing technologies become widely available and less costly, genomic testing attracts more attention, especially with many advantages (e.g. being reliable). Although the overall health benefits of genomic risk prediction have yet to be established, it is hoped that by incorporating into a patient's clinical information (e.g. electronic health records) and other resources, genomic risk prediction could attain improved accuracy, and ultimately can be used for improving clinical and health outcomes while reducing costs in the foreseeable future.

5 Conclusions

In summary, we proposed an ELR approach, Fam-ELR, for family-based genomic risk prediction. Fam-ELR not only incorporates a clustered ROC curve method to consider correlations among family samples, but also uses a computationally efficient tree-assembling procedure for variable selection and model building. Through simulations and applications on a family-based GWAS dataset of CD, we proved that Fam-ELR can incorporate genetic variants with small or moderate marginal effects and their interactions to form a risk prediction model with improved accuracy, and thus could be quite useful for high-dimensional family-based risk prediction especially on complex disease with limited knowledge of disease etiology.

Acknowledgements

The authors are thankful to Department of Surgery, Stanford University and Epidemiology and Biostatistics, Michigan State University, USA.

Compliance with ethics guidelines

Hui AN, Chang-shuai WEI, Oliver WANG, Da-hui WANG, Liang-wen XU, Qing LU, and Cheng-yin YE declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Abraham G, Inouye M, 2015. Genomic risk prediction of complex human disease and its clinical application. *Curr Opin Genet Dev*, 33:10-16. https://doi.org/10.1016/j.gde.2015.06.005

Anney RJL, Lasky-Su J, Ó'Dúshláine C, 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(8):1369-1378. https://doi.org/10.1002/ajmg.b.30871

Chatterjee N, Wheeler B, Sampson J, et al., 2013. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet*, 45(4):400-405. https://doi.org/10.1038/ng.2579

Choi S, Bae S, Park T, 2016. Risk prediction using genome-wide association studies on type 2 diabetes. *Genomics Inform*, 14(4):138-148. https://doi.org/10.5808/GI.2016.14.4.138

de los Campos G, Naya H, Gianola D, et al., 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1):375-385. https://doi.org/10.1534/genetics.109.101501

Ferreira MAR, O'Donovan MC, Meng YA, et al., 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet*, 40(9):1056-1058. https://doi.org/10.1038/ng.209

Ginsburg GS, Willard HF, 2009. Genomic and personalized medicine: foundations and applications. *Transl Res*, 154(6): 277-287. https://doi.org/10.1016/j.trsl.2009.09.005

Goes FS, Hamshere ML, Seifuddin F, et al., 2012. Genome-wide association of mood-incongruent psychotic bipolar disorder. *Transl Psychiatry*, 2(10):e180. https://doi.org/10.1038/tp.2012.106

Goes FS, McGrath J, Avramopoulos D, et al., 2015. Genome-wide association study of schizophrenia in Ashkenazi Jews. *Am J Med Genet B Neuropsychiatr Genet*, 168(8): 649-659. https://doi.org/10.1002/ajmg.b.32349

Janssens ACJW, van Duijn CM, 2008. Genome-based prediction of common diseases: advances and prospects. *Hum Mol Genet*, 17(R2):R166-R173. https://doi.org/10.1093/hmg/ddn250

Kazdin AE, 1997. Practitioner review: psychosocial treatments for conduct disorder in children. *J Child Psychol Psychiatry*, 38(2):161-178. https://doi.org/10.1111/j.1469-7610.1997.tb01851.x

Lasky-Su J, Neale BM, Franke B, et al., 2008. Genome-wide association scan of quantitative traits for attention deficit hyperactivity disorder identifies novel associations and confirms candidate gene associations. *Am J Med Genet B Neuropsychiatr Genet*, 147B(8):1345-1354. https://doi.org/10.1002/ajmg.b.30867

Maller J, George S, Purcell S, et al., 2006. Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat Genet*, 38(9):1055-1059. https://doi.org/10.1038/ng1873

Marchini J, Donnelly P, Cardon LR, 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 37(4):413-417. https://doi.org/10.1038/ng1537

Meigs JB, Shrader P, Sullivan LM, et al., 2008. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med*, 359(21):2208-2219. https://doi.org/10.1056/NEJMoa0804742

Need AC, Attix DK, McEvoy JM, et al., 2009. A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. *Hum Mol Genet*, 18(23):4650-4661. https://doi.org/10.1093/hmg/ddp413

Obuchowski NA, 1997. Nonparametric analysis of clustered ROC curve data. *Biometrics*, 53(2):567-578. https://doi.org/10.2307/2533958

Pappa I, St Pourcain B, Benke K, et al., 2016. A genome-wide approach to children's aggressive behavior: the EAGLE consortium. *Am J Med Genet B Neuropsychiatr Genet*, 171(5):562-572. https://doi.org/10.1002/ajmg.b.32333

Rietveld CA, Esko T, Davies G, et al., 2014. Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc Natl Acad Sci USA*, 111(38):13790-13794. https://doi.org/10.1073/pnas.1404623111

Sherva R, Wang Q, Kranzler H, et al., 2016. Genome-wide association study of cannabis dependence severity, novel risk variants, and shared genetic risks. *JAMA Psychiatry*, 73(5):472-480. https://doi.org/10.1001/jamapsychiatry.2016.0036

Shieh Y, Hu DL, Ma L, et al., 2016. Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Res Treat*, 159(3):513-525. https://doi.org/10.1007/s10549-016-3953-2

Smith JA, Ware EB, Middha P, et al., 2015. Current applications of genetic risk scores to cardiovascular outcomes and subclinical phenotypes. *Curr Epidemiol Rep*, 2(3): 180-190. https://doi.org/10.1007/s40471-015-0046-4

Sonuga-Barke EJS, Lasky-Su J, Neale BM, 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(8): 1359-1368. https://doi.org/10.1002/ajmg.b.30860

Wackerly DD, Mendenhall III W, Scheaffer RL, 2008. Mathematical Statistics with Applications, 7th Ed. Thomson, Belmont, CA, USA.

Wei CS, Anthony JC, Lu Q, 2012. Genome-environmental risk assessment of cocaine dependence. *Front Genet*, 3:83. https://doi.org/10.3389/fgene.2012.00083

Wei CS, Schaid DJ, Lu Q, 2013. Trees assembling Mann-Whitney approach for detecting genome-wide joint association among low-marginal-effect loci. *Genet Epidemiol*, 37(1):84-91.

https://doi.org/10.1002/gepi.21693

Wen YL, Burt A, Lu Q, 2017. Risk prediction modeling on family-based sequencing data using a random field method. *Genetics*, 207(1):63-73. https://doi.org/10.1534/genetics.117.199752

Wray NR, Lee SH, Mehta D, et al., 2014. Research review: polygenic methods and their application to psychiatric traits. *J Child Psychol Psychiatry*, 55(10):1068-1087. https://doi.org/10.1111/jcpp.12295

Yang J, Benyamin B, McEvoy BP, et al., 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565-569. https://doi.org/10.1038/ng.608

Ye C, Zhu J, Lu Q, 2011a. A clustered optimal ROC curve method for family-based genetic risk prediction. *Stat Interface*, 4(3):373-380. https://doi.org/10.4310/SII.2011.v4.n3.a11

Ye C, Cui Y, Wei C, et al., 2011b. A non-parametric method for building predictive genetic tests on high-dimensional data. *Hum Hered*, 71(3):161-170. https://doi.org/10.1159/000327299

List of electronic supplementary materials

Table S1 Significant interaction effects identified by logistic regression in the genome-wide prediction

中文概要

题 目：基于家系数据集群化似然比算法的疾病基因组遗传风险预测研究

目 的：作为遗传研究中最常用的设计之一，基于家系数据的实验设计因其优势而得到了广泛认可，例如家系数据在人群分层和混合情况下表现出来的稳健性。在疾病风险预测中，研究者对如何基于家系遗传数据，寻找和分析遗传标记的作用非常感兴趣。本研究旨在开发一种新的统计方法，用于基于家系数据的遗传风险预测。

创新点：期望新方法能够捕捉小或中等边际效应的遗传因子，及其相互作用，与基于家族史或家系数据的现有风险预测方法相比，具有更高的预测准确性。

方 法：在这项研究中，我们提出了集群化似然比（ELR）的新方法，Fam-ELR，用于家系数据的基因组疾病风险预测。Fam-ELR 采用集群化的受试者工作特征曲线（ROC）方法来考虑家系样本内部的相关性，并使用计算有效的集群树进行变量选择和模型构建。

结 论：通过模拟，Fam-ELR 显示了其在各种疾病遗传模型和谱系结构中的稳健性，并且获得了比现有的两种基于家系数据的风险预测方法更好的性能。同时，在基于全基因组行为障碍家系数据集的实际应用中，Fam-ELR 展示了其将潜在风险预测因子和其相互作用整合到模型中以提高准确性的能力，尤其是在全基因组水平上。通过比较现有方法，例如遗传风险评分方法等，Fam-ELR 被证实具有将较小或中等边际效应的遗传变异及其相互作用纳入改进的风险预测模型的能力。因此，它是一种强有力且实用的方法，适用于基于家系数据的高维度遗传风险预测中，特别是对于病因未知或知之甚少的人类复杂疾病。

关键词：家系数据研究；遗传风险预测；高维数据