Predicting the status of human complex diseases with random forest and polygenic risk scores

Jiawen Xu
Sichuan University West China Hospital
Jun Ma
Sichuan University West China Hospital
Yi Zeng
Sichuan University West China Hospital
Haibo Si
Sichuan University West China Hospital
Yuangang Wu
Sichuan University West China Hospital
Shaoyun Zhang
Sichuan University West China Hospital
Bin Shen (shenbin_1971@163.com)
Sichuan University West China Hospital

Research Article

Keywords: random forest, polygenic risk scores, disease prediction

Posted Date: May 20th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1631486/v1

Abstract

In recent years, polygenic risk score (PRS) analysis has become one of the most practical ways to leverage genome wide association studies (GWAS) findings for disease prediction. This approach is useful for addressing the challenge to translate the vast knowledge of complex disease genetics into clinically usable information. As machine learning is being widely applied to life science and PRS analysis comes into wide use for disease prediction, we systematacially evaluated the performance of random forest and PRS in predicting the status of complex diseases. Simulation studies were conducted by the GWAsimulator software, considering various genetic effects, genetic models and sample sizes. Two target complex disease related diseases and two environmental exposure factors were also simulated to obtain the additional genetic information of target complex disease, which were generally ignored in previous PRS studies. We found that PRS-based disease prediction using random forest had moderate accuracies (~ 70%) under various scenarios simulated by this study. The genetic effects of simulated disease loci showed the most significant impact on the performance of PRS-based disease prediction. This novel approach can leverage pleiotropy and gene-environment interactions. Furthermore, it is an attempt combining publicly available summary statistics and individual-level genotype data. We hope that this study provides useful information for further approaches development and disease prediction.

Key Points

• We evaluated the performance of random forest and PRS in predicting the status of complex diseases.

• Target complex disease related diseases and environmental exposure factors were taken into consideration for obtaining additional useful information.

• Various genetic effects, genetic models and sample sizes were simulated.

Introduction

Complex diseases, with a multifactorial etiology, are frequently encountered in health care. They are generally caused by multiple genes and environmental factors, involving gene–gene and/or gene–environment interactions. A plethora of susceptibility loci have been identified by genome wide association studies (GWAS) and follow-up meta-analyses for various common complex diseases[1]. Undoubtedly, these findings will result in a better understanding of the pathogenesis of complex diseases and facilitate the development of novel therapeutic options. Furthermore, leveraging these discoveries to better predict the status of complex diseases will greatly improve the prevention and treatment of complex diseases and attracts great interest recently[2]. As is well-known, a monogenic disease can be accurately predicted by the corresponding disease-causing mutation. However, although several approaches have been raised for complex disease prediction (such as penalized regression methods [3, 4]

and random-effects models [5]), it still remains a great challenge because of the complex genetic architecture.

Currently, one of the most practical ways to leverage recent GWAS findings for disease prediction is the polygenic risk score (PRS) analysis [6]. Typically, a PRS is calculated as the weighted sum of a number of high risk loci [7]. It combines the modest effects of multiple disease associated SNPs into a single variable, therefore has higher power than that of a single SNP. PRS have been created for many complex diseases, such as cardiovascular disease[8], multiple sclerosis [9] and schizophrenia [10]. They are used in different ways in recent years since more genetic data becomes readily available, such as Mendelian randomization studies[11] and disease prediction [12]. Among these disease prediction may be particularly useful for addressing the current challenge to translate the vast genetical knowledge of complex diseases into clinically usable information. For instance, several PRS have been proposed to optimize the use of genetic information of type 1 diabetes and ultimately improve its prediction and diagnosis [13]. Additionally, a previous study demonstrated that PRS was a powerful predictor for patients with first-episode psychosis using logistic regression [14].

But on the other side, using PRS analysis for disease prediction suffers the limitations from various factors nowadays. First, genetic correlations often exist among correlated diseases [15], which were hardly taken into consideration currently. Second, some environmental risk factors are themselves heritable (e.g. lipid fractions), and can mediate part of the genetic risk of the target disease. This was also mostly ignored in current PRS-based disease prediction. Taking into account target complex disease related diseases and environmental exposure factors may help obtain additional useful information for disease prediction.

As a subfield of computer science, machine learning algorithms play an essential role in the process of knowledge extraction [16, 17] and have been successfully applied in clinical field. For instance, Capper et al. used a machine learning approach to classify brain tumors on the basis of DNA methylation recently. Compared to standard methods, it resulted in a change of diagnosis in up to 12% of prospective cases[18]. Random forest is a tree based machine learning algorithm that consists of a collection of randomized decision trees[19]. Preliminary experiments showed that compared to several other popular machine learning algorithms (e.g. support vector machines), random forest achieved the highest accuracy [20].

Generally, a machine learning algorithm is used to train a classification model for separating samples of different classes (e.g. healthy or ill) based on variables (e.g. SNPs in a GWAS). In this circumstance, the whole original genomic data sets are generally used. However, instead of using these whole original genomic data, utilizing the combined genetic information of target disease associated genetic loci has the potential to enhance the performance of machine learning for disease prediction. To the best of our knowledge, the performance of combining PRS and machine learning for disease prediction remains largely unknown.

As machine learning and PRS are becoming more and more popular in genetic studies of complex diseases, we systematacially assessed the performance disease prediction with a combination of random forest and PRS. We first used random forest to train a classification model for separating samples of different health status (healthy or ill) based on PRS matrix and phenotypical data. Notably, PRS were constructed using the identified loci associated with target complex disease (e.g. type 2 diabetes), the complex diseases genetically related to target complex disease (e.g. hypertension, obesity) and environmental exposures (e.g. smoking, lack of exercises). We illustrated the feasibility and performance of this disease prediction approach through extensive genetic simulation. Our results may provide valuable information for applying random forest to PRS matrix for complex disease prediction.

Methods

1. Primary arithmetic steps

1.1 PRS matrix calculation

The independent SNP sets associated with target complex disease, the complex diseases related to target complex disease and environmental exposures can be derived from previous GWAS and used for PRS calculation in this study. Generally, genome wide significant loci (SNPs with GWAS *P* values < 5×10^{-8}) were chosen for analysis. Let E_i denotes the genetic effect parameter (beta or logarithm of odds ratio) of the i*th* SNP (i = 1,2,….n) associated with the k*th* (k = 1,2,….b) predictive factor (including target complex disease, the complex diseases related to target complex disease and environmental exposures). E_i is obtained from previously GWAS of complex diseases or environmental exposures. C_i denotes the risk allele dose of the i*th* SNP driven from individual level genotype data of target complex disease. For the u*th* individual (u = 1,2,….a), the weighted PRS of k*th* predictive factor can be calculated by:

$$PRS_{uk} = \sum_{i=1}^{n} E_i C_i$$

1.2 Applying random forest to PRS matrix

Let $X = \begin{bmatrix} GRS_{11} & \cdots & GRS_{1b} \\ \vdots & \ddots & \vdots \\ GRS_{a1} & \cdots & GRS_{ab} \end{bmatrix}$ denotes the PRS matrix of *b* predictive factors. *a* denotes the total number of study subjects with individual level genotype and phenotype data for target complex disease. Random forest (implemented by the 'randomForest' package of R software [21], http://www.r-project.org/) is then applied to PRS matrix *X* and disease phenotypes of the target disease to build a classifier for disease prediction. Briefly, the PRS matrix *X* and disease phenotypes of the target disease are split into training and test sample sets (e.g. 80% for training and 20% for test). Random forest is a tree based machine learning algorithm. Every decision tree in the random forest is built using a random subset of samples

and variables from the training sample set. The final classifier is an ensemble of many individual decision trees. It outputs the disease status that is predicted by the majority of those trees to classify the test sample. The performance of random forest for disease prediction is evaluated by the training accuracy and testing accuracy of the classifier, calculated by the 'randomForest' package.

2. Simulation studies

Simulation studies were conducted to evaluate the performance of predicting the risks of complex diseases with random forest and PRS. GWAsimulator software [22] (http://biostat.mc.vanderbilt.edu/GWAsimulator) was used for simulating individual level genotype and phenotype data. GWAsimulator implements a rapid moving-window algorithm and can simulate genotype and phenotype data based on real Illumina HumanHap550 chip data [22]. As disease phenotypes are generally determined by multiple factors and complex diseases generally result from the interaction of genes and environmental factors. Without loss of generality, we simulated one target complex disease, two complex diseases related to the target complex disease, and two environmental exposure factors for PRS calculation. 15 disease loci were simulated for every human disease and environmental exposure, each was on a different chromosome. 5000 SNP loci were simulated for each chromosome. The population prevalence was set to 0.1 in GWAsimulator. Using GWAsimulator, we generated various scenarios, including relative risks of causal loci, inheritance models, and sample sizes (Table 1). Following the standard approach, the PRS matrix was calculated by PLINK [23]. Random forest was implemented by the 'randomForest' package [21] of R (http://www.r-project.org/). We used 80% of the simulated samples to train the random forest classifier. The disease states of the remaining 20% of the simulated samples were predicted by the generated random forest classifier. 1000 simulations were conducted for each parameter setting. The training accuracy and testing accuracy for disease status prediction were reported by the 'randomForest' package, respectively.

Table 1
Parameter settings of simulation studies

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Relative risks of causal loci	1.1	1.5	*2.0*	2.5	3.0
Inheritance models[a]	D	*M*	R	-	-
Sample sizes	2000	4000	6000	*8000*	10000

Note: [a] D: dominance model; M: multiplicative model; R: recessive model; [b] The default parameters are highlighted in italic; [c] For parameters with less than 5 settings, the blanks are filled by dashes.

Result

Table 2 summarizes the simulation results of various simulation parameters. The relative risks of simulated causal loci showed significant impact on the prediction accuracy. The prediction accuracy

increased with the increasing of relative risks. The testing accuracy was 0.506 when OR=1.1 for simulated causal loci, and achieved 0.705 when OR 3.0 (Fig. 1A). Under the various inheritance models simulated by this study, we observed the highest testing accuracy (0.654) for multiplicative model. The testing accuracies were basically identical for dominance model (0.582) and recessive model (0.584) (Fig. 1B). As shown by Fig. 1C, the testing accuracy appeared to slightly increase with increased sample sizes in this study. The testing accuracy was 0.642 when sample size equaled 2000 and achieved 0.655 when sample size increased to 10000.

Table 2
Simulation results of all parameter settings

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Relative risks of causal loci	**1.1**	**1.5**	2.0	**2.5**	**3.0**
Training accuracy	0.504	0.586	0.652	0.688	0.705
Testing accuracy	0.506	0.586	0.654	0.689	0.705
Inheritance models[a]	**D**	M	**R**	**-**	**-**
Training accuracy	0.582	0.652	0.585	-	-
Testing accuracy	0.582	0.654	0.584	-	-
Sample sizes	**2000**	**4000**	**6000**	8000	**10000**
Training accuracy	0.640	0.649	0.652	0.652	0.654
Testing accuracy	0.642	0.651	0.653	0.654	0.655

Note: [a] D: dominance model; M: multiplicative model; R: recessive model; [b] The default parameters are highlighted in italic; [c] For parameters with less than 5 settings, the blanks are filled by dashes.

Discussion

The prediction of diseases is one of the most interesting and challenging tasks in the genetic researches of complex diseases, as it can facilitate the subsequent clinical management of patients. Random forest has been widely used in life science recently and PRS analysis is an increasingly popular method for utilizing GWAS summary data, which has been used for disease prediction[12]. To the best of our knowledge, no research has applied random forest to PRS matrix and disease phenotypes for prediction of disease status. Furthermore, previous PRS-based disease prediction generally did not consider the genetic information of the diseases and environmental exposures related to target disease, which should provide more useful genetic information for disease risk prediction. We illustrated the feasibility and performance of applying random forest to PRS matrix of target complex disease, the complex diseases related to target complex disease and environmental exposures for the prediction of disease status in this study through simulation analyses.

We presented the first comprehensive evaluation of applying random forest to PRS matrix for the prediction of disease state, including an assessment of the effects of changes in various parameters. We observed that the relative risks of causal loci had greater influence on the prediction accuracy. Under various relative risks of causal loci, the testing accuracy varied from 0.506 to 0.705. The training and testing accuracies both achieved 0.705 (the maximum value in this study) when relative risks of causal loci increased to 3.0. Multiplicative model had the highest testing accuracy (0.654) among the various inheritance models simulated by this study. Sample sizes showed limited influence on prediction accuracy. The potential explanation for the results of relative risks is that when causal loci have stronger causal correlations with a certain complex disease or environmental exposure, they take a greater part of the disease risk factors and can therefore better predict the target disease. The same goes with the results of inheritance models. When relative risks of causal loci are greater than 1, as in this study, multiplicative model obtains the greatest overall effect, and thus has the best prediction accuracy.

This study has several innovations. First, this novel approach can leverage pleiotropy and gene-environment interactions, which were generally ignored in previous PRS-based disease prediction. Better prediction accuracy should be obtained after considering the genetic information of all target diseases related factors. Second, random forest itself has the following distinct advantages. Although we only considered genetic factors in our simulation studies and the absence of other influential factors may affect the prediction accuracies as non-genetic factors generally have great influences on complex diseases. It's easy to join other information features (e.g. clinical outcomes, related physiological parameters and living habits) to random forest, which may further enhance the prediction accuracy. Additionally, random forest has the potential to unravel interactions among variables, which are ubiquitous in life science related data sets. Interactions can for example exist among SNPs in GWAS [24], among cellular levels of gene-products in gene-expression studies [25]. Furthermore, preliminary experiments showed that random forest achieved the highest accuracy compared to several other popular machine learning algorithms (e.g. support vector machines) [20]. Third, this approach is an attempt to combine publicly available summary statistics and individual-level genotype data.

Machine learning methods have become a popular tool for medical researchers and have been widely applied to disease prognosis and prediction [26–28]. Concerning the future of disease prediction, novel creative researches should be conducted in the subsequent studies. For instance, researchers can include large amounts of omics data (e.g. DNA methylation, gene expression profile) for disease prediction, which may provide additional useful information. Furthermore, prediction analyses can also be conducted across ethnically diverse data to predict disease status for individuals from various races.

There are several issues that should be noted. First, we used the default values in random forest in our study. A previous research thoroughly examined the effects of changes in the parameters of random forest (specifically mtry, ntree, nodesize). The results demonstrated that changes in these parameters have in most cases negligible effects, suggesting that the default values are often good options[25]. Second, for practical use, preliminary analysis (such as LDSC analysis) should be conducted in the first place for selecting the target complex disease related diseases and environmental exposure factors. Then

our prediction approach can be applied to all these identified risk factors. Third, in addition to random forest, there are multiple other optional machine learning techniques which have successfully been used for disease prediction (e.g. artificial neural networks[29] and support vector machines[30]).

Conclusion

In conclusion, we conducted an assessment of predicting the status of complex diseases with random forest and PRS. Simulation studies demonstrated that this approach had moderate training and testing accuracies for disease prediction. Given that machine learning is increasingly popular, we hope that our research will provide additional insight into related areas and enlighten further studies.

Declarations

Author Contribution Statements

(I) Conception and design: Author 1, Author 7

(II) Administrative support: Author 2, Author 7

(III) Provision of study materials: Author 1, Author 3, Author 4

(IV) Collection and assembly of data: Author 1, Author 4

(V) Data analysis and interpretation: Author 1, Author 5, Author 6

(VI) Manuscript writing: Author 1, Author 2

(VII) Final approval of manuscript: All authors

Availability of Data and Material Statement

Random forest was implemented by the 'randomForest' package of R (http://www.r-project.org/). GWAsimulator software (http://biostat.mc.vanderbilt.edu/GWAsimulator) was used for simulating individual level genotype and phenotype data.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (grant number 81974347 and 81802210); the Department of Science and Technology of Sichuan Province (grant number 2021YFS0122). Financial support had no impact on the outcomes of this study.

Acknowledgements

Not applicable

References

1. Welter D, Macarthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L: **The NHGRI GWAS catalog, a curated resource of sNP-trait associations**. Nucleic Acids Research 2014, **42**(Database issue):1001–1006.
2. Manolio TA: **Bringing genome-wide association findings into clinical use**. Nature Reviews Genetics 2013, **14**(8):549–558.
3. Wu TT, Chen YF, Hastie T, Sobel E, Lange K: **Genome-wide association analysis by lasso penalized logistic regression**. Bioinformatics 2009, **25**(6):714.
4. Won S, Choi H, Park S, Lee J, Park C, Kwon S: **Evaluation of Penalized and Nonpenalized Methods for Disease Prediction with Large-Scale Genetic Data**. Biomed Res Int 2015, **2015**(1):605891.
5. Speed D, Balding DJ: **MultiBLUP: improved SNP-based prediction for complex traits**. Genome Research 2014, **24**(9):1550–1557.
6. Amin N, Duijn CMV, Janssens ACJW: **Genetic Scoring Analysis: a way forward in Genome Wide Association Studies?** European Journal of Epidemiology 2009, **24**(10):585–587.
7. Frank D: **Correction: Power and Predictive Accuracy of Polygenic Risk Scores**. Plos Genetics 2013, **9**(4):e1003348.
8. Thanassoulis G, Peloso GM, Pencina MJ, Hoffmann U, Fox CS, Cupples LA, Levy D, D'Agostino RB, Hwang SJ, O'Donnell CJ: **A Genetic Risk Score Is Associated With Incident Cardiovascular Disease and Coronary Artery CalciumClinical Perspective**. Circulation Cardiovascular Genetics 2012, **5**(5):113–121.
9. De Jager PL, Chibnik LB, Cui J, Reischl J, Lehr S, Simon KC, Aubin C, Bauer D, Heubach JF, Sandbrink R: **Integration of genetic risk factors into a clinical algorithm for multiple sclerosis susceptibility: a weighted genetic risk score**. Lancet Neurology 2009, **8**(12):1111–1119.
10. 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**(7256):748.
11. Palmer TM, Lawlor DA, Harbord RM, Sheehan NA, Tobias JH, Timpson NJ, Davey SG, Sterne Jonathan AC: **Using multiple genetic variants as instrumental variables for modifiable risk factors**.

Statistical Methods in Medical Research 2012, **21**(3):223.

12. Weijmans M, de Bakker PI, Van dGY, Asselbergs FW, Algra A, Jan dBG, Spiering W, Visseren FL: **Incremental value of a genetic risk score for the prediction of new vascular events in patients with clinically manifest vascular disease**. Atherosclerosis 2015, **239**(2):451–458.
13. Redondo MJ, Oram RA, Steck AK: **Genetic Risk Scores for Type 1 Diabetes Prediction and Diagnosis**. Curr Diab Rep 2017, **17**(12):129.
14. Vassos E, Forti MD, Coleman J, Iyegbe C, Prata D, Euesden J, O'Reilly P, Curtis C, Kolliakou A, Patel H: **An Examination of Polygenic Score Risk Prediction in Individuals With First-Episode Psychosis**. Biological Psychiatry 2016, **81**(6):470–477.
15. Ohn JH: **The landscape of genetic susceptibility correlations among diseases and traits**. J Am Med Inform Assoc 2017, **24**(5):921–926.
16. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A: **Machine learning in bioinformatics**. Briefings in Bioinformatics 2006, **7**(7):86–112.
17. Verikas A, Gelzinis A, Bacauskiene M: **Mining data with random forests: A survey and results of new tests**. Pattern Recognition 2011, **44**(2):330–349.
18. Capper D, Jones D, Sill M, Hovestadt V, Schrimpf D, Sturm D, Koelsche C, Sahm F, Chavez L, Reuss DE: **DNA methylation-based classification of central nervous system tumours**. Nature 2018, **555**(7697).
19. Breiman L: **Random Forests**. Machine Learning 2001, **45**(1):5–32.
20. Ellis K, Godbole S, Marshall S, Lanckriet G, Staudenmayer J, Kerr J: **Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms**. Frontiers in Public Health 2014, **2**:36.
21. Liaw A, Wiener M: **Classification and regression by randomForest**. **R News 2**:**18**–**22**. 2001, **23**.
22. Li C, Li M: **GWAsimulator: a rapid whole-genome simulation program**. Bioinformatics 2008, **24**(1):140–142.
23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, Bakker PIWD, Daly MJ: **PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses**. American Journal of Human Genetics 2007, **81**(3):559–575.
24. Moore JH, Asselbergs FW, Williams SM: **Bioinformatics challenges for genome-wide association studies**. Bioinformatics 2010, **26**(4):445–455.
25. Díaz-Uriarte R, Andrés SAD: **Gene selection and classification of microarray data using random forest**. Bmc Bioinformatics 2006, **7**(1):3.
26. Cruz JA, Wishart DS: **Applications of Machine Learning in Cancer Prediction and Prognosis**. Cancer Informatics 2006, **2**(1):59–77.
27. Cicchetti DV: **Neural networks and diagnosis in the clinical laboratory: state of the art**. Clinical Chemistry 1992, **38**(1):9–10.

28. Kononenko I: **Machine learning for medical diagnosis: history, state of the art and perspective**. Artificial Intelligence in Medicine 2001, **23**(1):89–109.
29. Bottaci L, Drew PJ, Hartley JE, Hadfield MB, Farouk R, Lee PW, Macintyre IM, Duthie GS, Monson JR: **Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions**. Lancet 1997, **350**(9076):469–472.
30. Akay MF: **Support vector machines combined with feature selection for breast cancer diagnosis**. Expert Systems with Applications 2009, **36**(2):3240–3247.

Figures

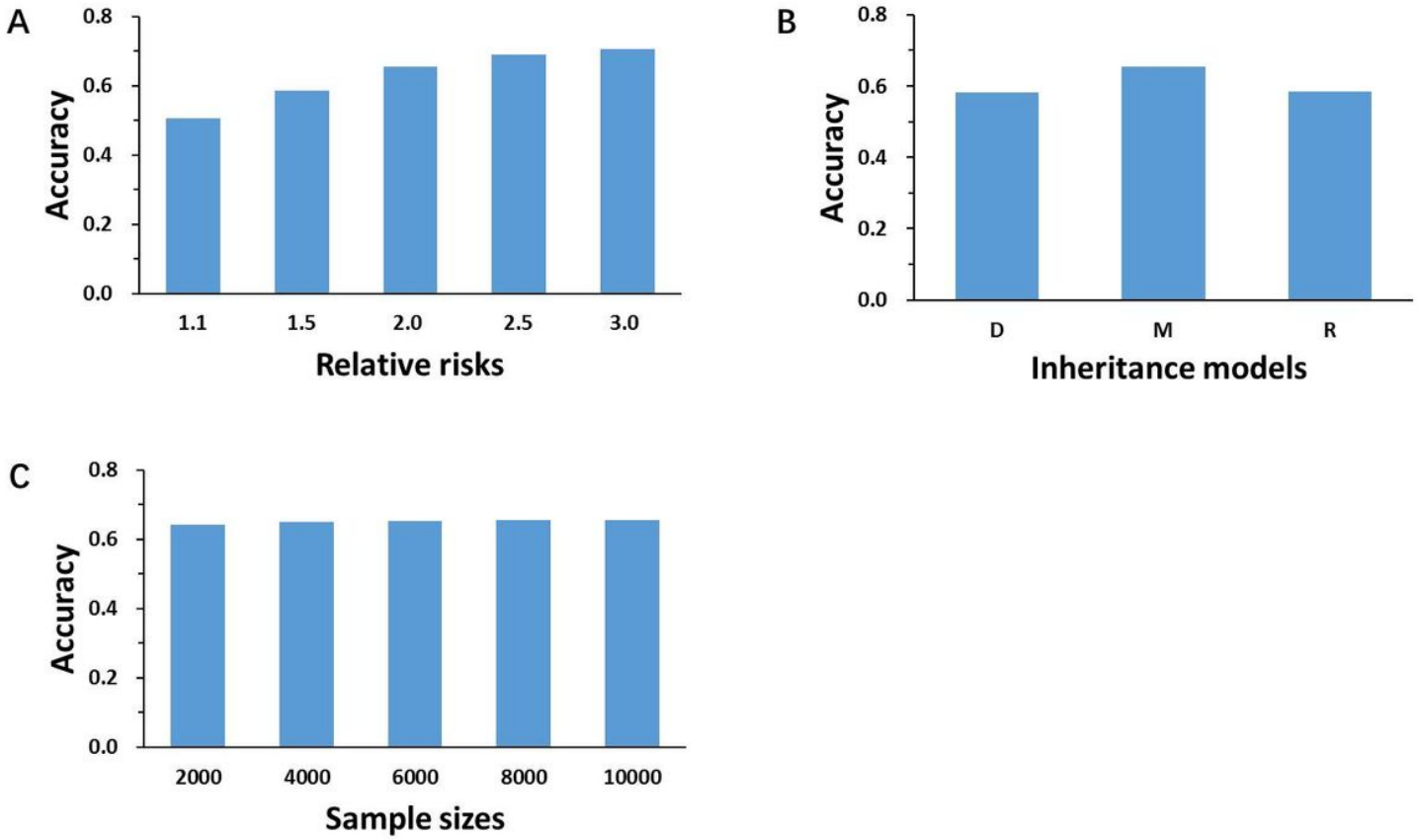

Figure 1

The simulation study results of relative risks of simulated causal loci, inheritance models and sample sizes. D: dominance model; M: multiplicative model; R: recessive model.