

Title

A Deep Learning-based Genome-wide Polygenic Risk Score for Common Diseases Identifies Individuals with Risk

Authors

Jiajie Peng,^{1,2*} Jingyi Li,^{1,2} Ruijiang Han,^{1,2} Yuxian Wang,^{1,2} Lu Han,^{1,2} Jinghao Peng,^{1,2} Tao Wang,^{1,2} Jianye Hao,³ Xuequn Shang,^{1,2} Zhongyu Wei^{4*}

Affiliations

¹School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China.

²Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an, 710072, China.

³College of Intelligence and Computing, Tianjin University, Tianjin, 300072, China.

⁴School of Data Science, Fudan University, Shanghai, 200433, China.

*Corresponding author. Email: jiajiepeng@nwpu.edu.cn or zywei@fudan.edu.cn

Keywords: disease risk prediction, deep learning, polygenic risk score

Abstract

Identifying individuals at high risk in the population is a key public health need. For many common diseases, individual susceptibility may be influenced by genetic variation. Recently, the clinical potential of polygenic risk score (PRS) has attracted widespread attention. However, the performance of traditional methods is limited in fitting capabilities of the linear model and unable to capture the interaction information between single nucleotide polymorphisms (SNPs). To fill this gap, a novel deep-learning-based model named DeepPRS is developed for scoring the risk of common diseases with genome-wide genotype data. Using the UK Biobank dataset, the evaluation shows that DeepPRS performs better than the other two existing state-of-art methods on Alzheimer's disease, inflammatory bowel disease, type 2 diabetes and breast cancer. Since DeepPRS does not only rely on the additive effect of risk

SNPs, DeepPRS has the chance to identify high-risk individuals even with few known risk SNPs.

1. Introduction

One of the main public health needs is to stratify the individuals for a given disease to identify individuals at high risk and enable enhanced screening or preventive therapies^[1]. For most human diseases, individual susceptibility is influenced by genetic variation to some extent^[2]. Consequently, one important approach for identifying individuals at high risk is to stratify individuals based on inherited DNA variation^[1]. According to the number of genes that cause the disease, the diseases involving genetic factors are traditionally divided into single-gene mendelian diseases and complex or common diseases^[3]. In the 1980s and 1990s, based on linkage analysis and fine mapping within large multiplex pedigrees, efforts to map disease genes mainly focused on rare diseases, monogenic diseases and syndrome-type diseases^[1]. By 2000, around 1000 single-gene inherited diseases had been characterized, including many diseases that have a significant impact on biomedicine, such as Huntington's disease^{[4][5]} and cystic fibrosis^{[6][7]}. However, linkage analysis is very limited for common, later-onset traits with more complex diseases, such as asthma, diabetes and depression^[8]. Until 2005, genome-wide association studies (GWAS) have identified a large number of genetic variants, mostly single nucleotide polymorphisms (SNPs)^[9]. The emergence of GWAS provided important clues for the discovery of genetic characteristics that influence the occurrence of complex diseases^[3].

In the decades since the first GWAS^[10], people's understanding of the genetic basis of common human diseases has been changed. For complex or common diseases, genetic susceptibility is jointly determined by thousands of mostly common variants, and a single variant has little effect on population risk^{[11][12]}. On the basis of GWAS research,

polygenic risk score (PRS), which quantifies individual genetic risk, has become a powerful tool for common disease risk prediction^[13]. PRS quantifies the degree of individual susceptibility to disease by calculating the cumulative effect of multiple susceptibility sites^[14]. The development of robust PRS for several common diseases has been catalyzed by the continuous expansion of GWAS dataset scale and the establishment of large-scale biobank supporting score verification^{[1][2][15][16]}. Many studies have demonstrated the utility of PRS for disease risk stratification as well as their implications for early disease detection, prevention, therapeutic intervention and life planning^[17]. The traditional PRS method quantifies the impact of variations on individuals based on a simple linear additive model. However, a popular thought suggests that the mechanisms of gene action and the structure of complex traits are actually much more complex than the additive model describes^[18]. For example, epistasis, defined in functional terms as an event where the influence of one locus depends on the genotype of another, is a type of non-additive genetic variation^[19]. A study shows that there is an interaction between HLA-C and ERAP1 for psoriasis, and ERAP1 variants only influence psoriasis susceptibility in individuals carrying the HLA-C risk allele^[20]. Therefore, the traditional additive method of PRS calculation not only fails to utilize the location information of variations, but also ignores the interaction information among various variations. In addition, the traditional PRS model is constructed using classifiers such as logistic regression model^[21] and penalty regression model^[22]. These models have limited model fitting capabilities, which also affects the accuracy of the final prediction results to a certain extent.

Recently, deep learning, a subfield of machine learning, has been successfully applied to several areas, such as medical imaging, health record natural language processing and generalized deep-learning methods for genomics^{[23][24][25][26]}.

Inspired by the success of deep learning in healthcare area, we hypothesize that deep learning can further enhance the predictive ability of PRS models by integrating large scale genotype data. In this study, we propose a deep learning-based approach to improve the traditional PRS model, and adopt bidirectional long short-term memory network (BiLSTM) as classification model. An efficient disease risk prediction method is presented, named DeepPRS, which can be used to calculate an individual's polygenic risk score and stratify the population. Here are our three contributions:

- We propose a deep neural network-based model, named DeepPRS, to generate polygenic risk score. Compared with the linear addition of the traditional method, we not only consider the position relationship of SNPs, but also capture the interaction information between long-distance genes on each chromosome through BiLSTM.
- Our risk analysis demonstrated that the disease risk of people with high deep polygenic risk score is much higher than people with low deep polygenic risk score. Therefore, DeepPRS can be used as an early warning sign to initiate disease prevention and screen in high-risk populations.
- Compared with traditional methods, DeepPRS needs less prior information. The performance of DeepPRS is better than the traditional methods, which needs the effect estimation given by GWAS as the genotype weight.

2. Result

2.1. Overview of the DeepPRS model

DeepPRS combines the genotype data^[27], summary statistics results of GWAS^{[28][29][30][31]}, and 1000 Genomes reference panel^[32] to predict the risk of common diseases (**Figure. 1a, Supplementary Figure. S1 and Table 1**). In short, we first carry out quality control (QC) on genotype dataset, and then a set of parameters based on GWAS results (P values) and linkage disequilibrium reference panel from 1000 Genomes of 503 Europeans (r^2) are applied to select SNPs for the downstream analysis. The feature selection of this step provides high-quality and low-dimensional input for generating deep polygenic risk score, which can greatly reduce the number of parameters and running time of the neural networks. Then, a novel encoding schema is proposed to represent genotype information. The two dimensions of the feature vector represents risk alleles (Alt allele) and non-risk alleles (Ref allele) respectively. Compared with the traditional additive encoding method, our encoding method can not only express the number of alleles carried, but also avoid additional quantitative relationships, and better deal with missing genotypes (**Figure. 1b**). Finally, the features are fed to deep neural networks for disease risk prediction. We first use a partial connected layer to connect SNPs to genes within 250 kilobases region according to the location information of SNPs, which greatly reduces the feature dimension and effectively reduces the over-fitting phenomenon. Next, we use the BiLSTM layer to capture the interaction information between long-distance genes from the forward and

backward directions at the same time, and finally we use the fully connected layer as the classifier. The prediction probability is used as the deep learning-based polygenic risk score (**Figure. 1c**).

We have studied four common diseases, including Alzheimer's disease (AD), inflammatory bowel disease (IBD), type 2 diabetes (T2D) and breast cancer (BC). The basic information of the dataset is shown in Table 1. The Area Under the Receiver Operating Characteristic Curve (AUC)^[33] of deep polygenic risk score is 0.7245, 0.6517, 0.6508 and 0.6227 for Alzheimer's disease, inflammatory bowel disease, type 2 diabetes and breast cancer in the test set, respectively (**Table 1**). The deep polygenic risk score with clinical features also performs well in the test dataset, with AUC of 0.8624, 0.6585, 0.7316 and 0.6660 for Alzheimer's disease, inflammatory bowel disease, type 2 diabetes and breast cancer, respectively (**Table 1**). Details of DeepPRS are provided in the Methods section.

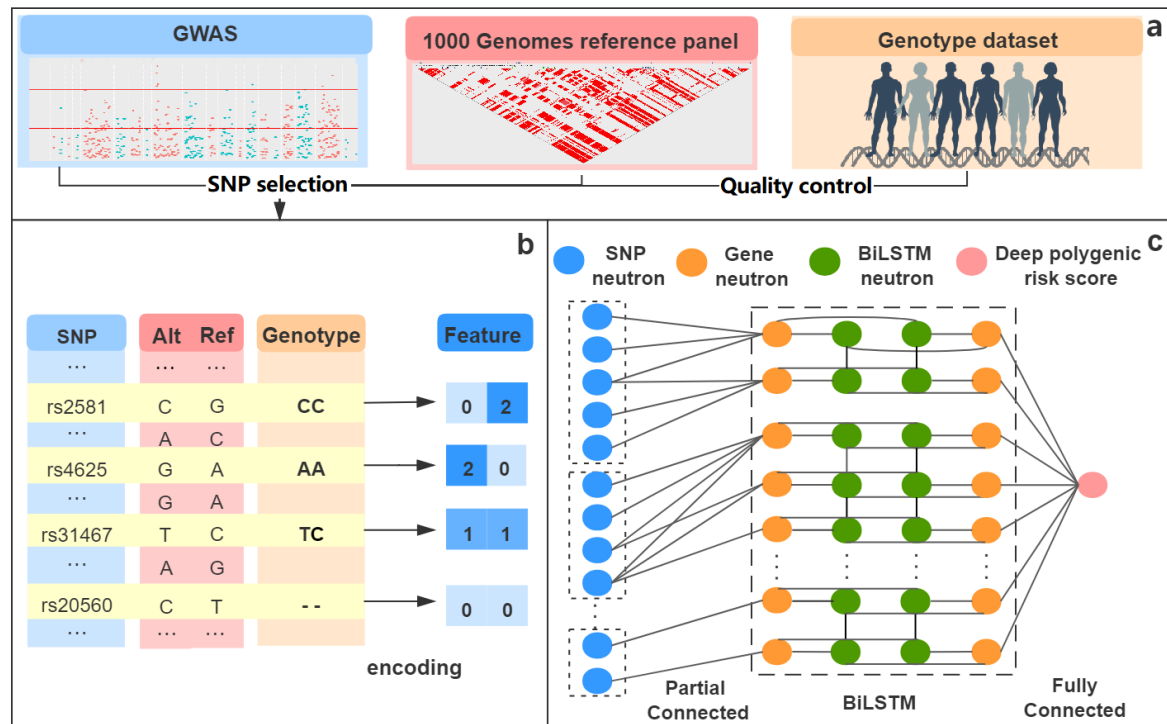


Figure. 1. Overview of the DeepPRS method. DeepPRS includes three main components. (a), SNP selection step considers the significance SNPs based on the GWAS result, the linkage disequilibrium reference panel from 1000 Genomes Europeans and the availability of genotype data. (b), The genotype information is encoded into two dimensions of feature vectors, where the first dimension represents the number of non-risk allele (Ref allele), the second dimension represents the number of risk allele (Alt allele), the value of both dimensions of missing genotype is set to zero. (c), We first groups SNP neurons according to chromosomes, then adopt a partial connected layer to connect SNP feature vectors to corresponding genes and use BiLSTM to capture the interaction relationship between genes. The deep polygenic risk score is obtained based on the fully connected layer with a sigmoid function.

Table 1. Deep polygenic risk score derivation and testing for four common diseases. Clinical features include age, sex, genotype measurement batch, genotype array, region of assessment center, Townsend Deprivation index at recruitment, education-qualifications, the first four of genetic principal components. The breast cancer analysis is restricted to female participants.

Disease	Discovery	Number of	DeepPRS	DeepPRS	with
	GWAS (n)	SNPs	AUC	clinical	features
				AUC	
Alzheimer's	17,008 cases;	771	0.7245	0.8624	
disease ^[28]	37,154 controls				
Inflammatory	12,882 cases;	2481	0.6517	0.6585	
bowel disease ^[29]	21,770 controls				
Type 2 diabetes ^[30]	26,676 cases;	5968	0.6508	0.7316	
	132,532 controls				
Breast cancer ^[31]	122,977 cases;	3830	0.6227	0.6660	
	105,974 controls				

2.2. DeepPRS can identify individuals with high risk for AD, IBD, T2D and BC

We perform a series of risk analyses to test whether DeepPRS can identify individuals with high disease risk. Similar to the method used in previous study^[1], a given threshold is used to separate the individuals based on deep polygenic risk score. The odds ratio (OR) is used to measure the association between individuals with high PRS versus remainders of population, and the percentile score is used to measure the location of

data (see details in the Methods section). Taking IBD as an example, the deep polygenic risk score of IBD is normally distributed across all tested individuals. We find that 4.19% of the population have inherited a genetic predisposition with more than threefold increased risk for IBD. Furthermore, 0.9% of the population have more than fourfold increased risk for IBD and 0.29% have more than fivefold increased risk (**Figure. 2a** and **Table 2**). The median DeepPRS_{IBD} percentile score is 69 for individuals with IBD, which is much higher than the 49 for non-IBD individuals (**Figure. 2b**). The dataset is divided into 100 groups according to the percentile of deep polygenic risk score. The risk of IBD increases sharply in the right tail of the deep polygenic score distribution, from 0.64% in the lowest percentile to 6.24% in the highest percentile, which shows the good risk prediction ability of DeepPRS (**Figure. 2c**). Similar results can also be found on AD, BC and T2D (**Supplementary Figure. S2-4**). The proportion of the population with OR greater than five are 7.90%, 0.06% and 0.15% for AD, BC and T2D, respectively (**Supplementary Figure. S2a-4a**). The median DeepPRS percentile scores are 81, 65, 67 for individuals with AD, BC and T2D respectively, which is much higher than the 49, 47, 48 for non-AD, non-BC and non-T2D individuals (**Supplementary Figure. S2b-4b**). The risks of AD, BC and T2D all increase sharply in the right tail of the deep polygenic score distribution, from 0.17%, 2.85%, 2.31% in the lowest percentile to 4.70%, 23.13%, 26.10% in the highest percentile respectively (**Supplementary Figure. S2c-4c**). Furthermore, DeepPRS performs substantially better than the lasso model for different given thresholds of defining high risk individuals (**Supplementary Table S1**). For top 1% of population versus the others,

using the deep polygenic risk score with genotype information only, the ORs of AD, IBD, T2D and BC are 8.97, 3.75, 3.70 and 3.29 respectively. For the lasso model, the ORs are much lower, which are 3.58, 2.15, 0.71 and 1.68 for AD, IBD, T2D and BC respectively.

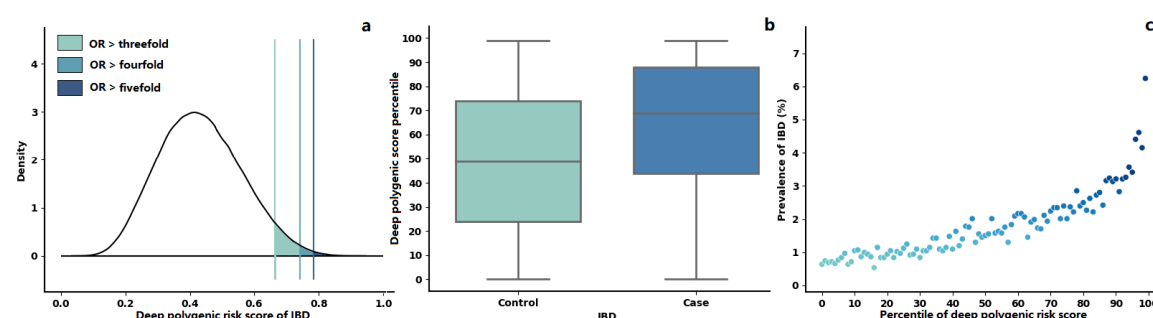


Figure. 2. Risk analysis for IBD based on DeepPRS. (a), Distribution of deep polygenic risk score of IBD in the UK Biobank dataset. The x axis represents deep polygenic risk score of IBD. Shading reflects the proportion of the population with three-, four-, and fivefold increased risk versus the remainder of the population. (b), Deep polygenic risk score percentile among IBD cases versus controls in the UK Biobank dataset. In each boxplot, the horizontal lines reflect the median, the top and bottom of each box reflect the quartile range, the whiskers reflect the maximum and minimum values within each group. (c), Prevalence of IBD according to 100 groups of the dataset binned according to the percentile of the deep polygenic risk score of IBD.

Table 2. Proportion of the population at three-, four- and fivefold increased risk for Alzheimer's disease, inflammatory bowel disease, type 2 diabetes and breast cancer.

High PRS definition	Individuals in UKB dataset (n)	% of individuals
Odds ratio ≥ 3.0		
Alzheimer's disease	243,567/351,022	69.39
Inflammatory bowel disease	16,457/392,613	4.19
Type 2 diabetes	14,190/346,333	4.10

Breast cancer	2,731/172,160	1.59
Odds ratio ≥ 4.0		
Alzheimer's disease	140,151/351,022	39.93
Inflammatory bowel disease	3,515/392,613	0.90
Type 2 diabetes	2,283/346,333	0.66
Breast cancer	328/172,160	0.19
Odds ratio ≥ 5.0		
Alzheimer's disease	27,706/351,022	7.90
Inflammatory bowel disease	1,138/392,613	0.29
Type 2 diabetes	532/346,333	0.15
Breast cancer	103/172,160	0.06

2.3. DeepPRS performs better than other two existing methods on four diseases

We study four common diseases, including AD, IBD, T2D and BC on UKB dataset. UKB dataset is divided into case group and control group according to the specific disease (**Table 1** and **Supplementary Figure. S5-8**). To avoid logic circle, we use the GWAS results that do not include UKB dataset. Taking AD as an example, the SNP selection is based on a recent GWAS involving 54,162 participants and a linkage disequilibrium reference panel of 503 Europeans from 1000 Genomes (**Table 1**). The UK Biobank dataset after quality control includes 351,022 participants, of which 2,066 are diagnosed with AD (**Table 1** and **Supplementary Figure. S5**). When $P < 0.0005$ and $r^2 < 1$, the prediction performance is the best, including 771 variants (**Table 1**). On

the task of disease risk prediction, we use ten-fold cross-validation^[34] for performance comparison (**Table 1** and **Figure. 3**). We predict the disease risk based on genotype only and genotype with clinical features using three methods, including DeepPRS, pruning and thresholding method^[21], and lasso model^[22] (**Supplementary Table S2-9**). Compared with the other two state-of-art methods, the results show that the DeepPRS achieves the best performance using genotype only or genotype with clinical features as input on all four diseases. The results also show that DeepPRS is not sensitive to the linkage disequilibrium parameters (**Supplementary Table S2-9**), indicating that DeepPRS can automatically deal with the redundancy information between different

SNPs without adding more prior information for linkage disequilibrium filtering.

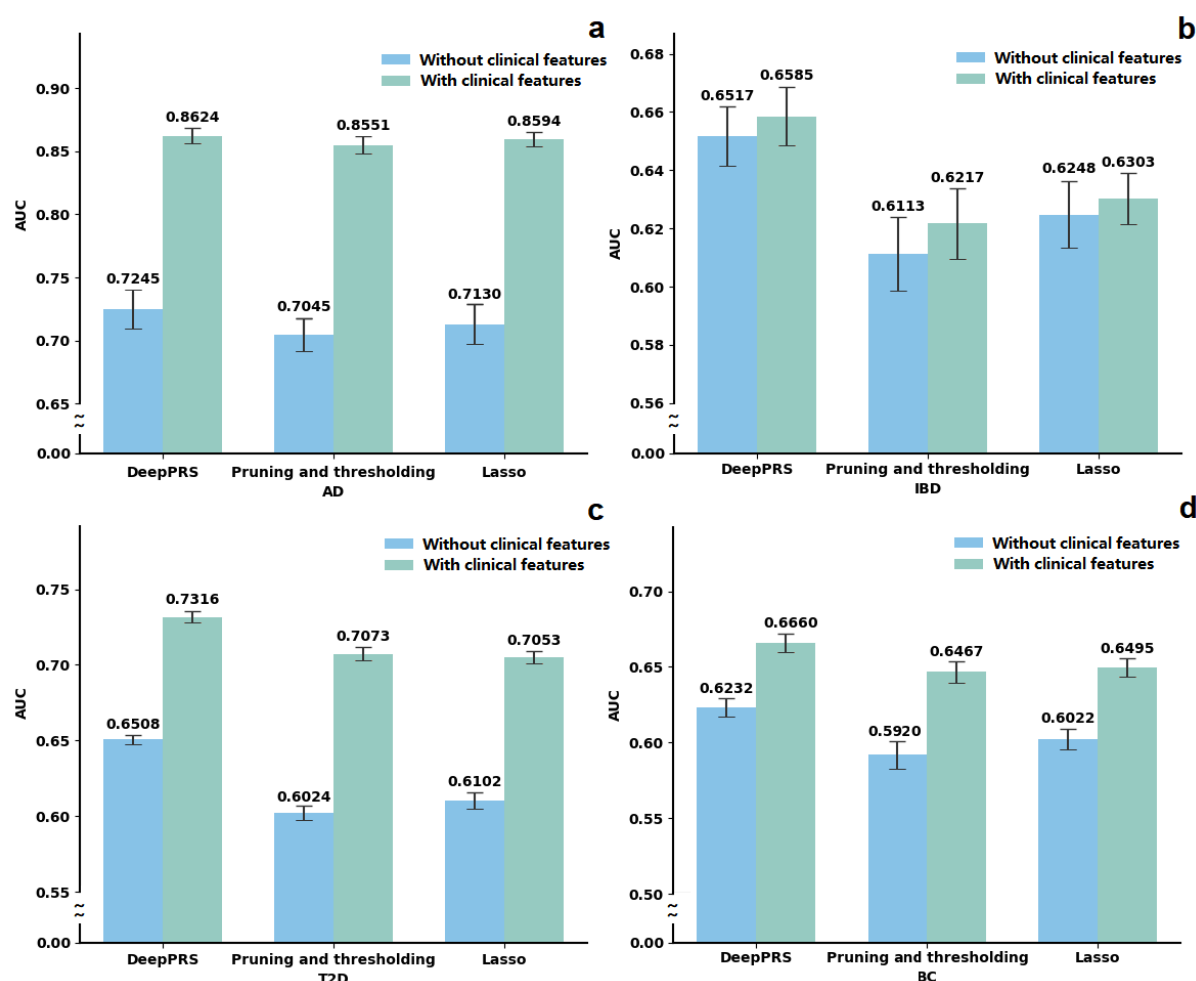


Figure 3. The results of risk prediction using genotype only and genotype with clinical features for Alzheimer's disease (a), inflammatory bowel disease (b), type 2 diabetes (c) and breast cancer (d). Compared with the pruning and thresholding method, and lasso model, the results show that the DeepPRS achieves the best performance on all the four diseases.

2.4. DeepPRS can identify the cases with few known risk SNPs

The traditional weighted-sum methods for calculating PRS are mainly based on the number of risk SNPs and their effect estimate (β) from GWAS result. Comparing with previous method, DeepPRS can find cases (diagnosed individuals) with few known risk SNPs. Taking AD as an example, three individuals in the case group only carry few risk

SNPs. Not surprisingly, their PRS percentile scores based on pruning and thresholding method are very low, which are all 1. However, their DeepPRS percentile scores are all 99, which are larger than threshold of the 9.53-fold risk (**Supplementary Table S10**). For IBD, PRS percentile scores based on pruning and thresholding method of three individuals are 5, 5 and 6, respectively. However, their DeepPRS percentile scores are all 99, which are larger than threshold of the 4.23-fold risk (**Supplementary Table S10**). For T2D, PRS percentile scores based on pruning and thresholding method of three individuals are 1, 2 and 9, respectively. However, their DeepPRS percentile scores are 75, 98 and 86, which are larger than threshold of the 2.34-fold risk (**Supplementary Table S10**). For BC, PRS percentile scores based on pruning and thresholding method of three individuals are 5, 19 and 9, respectively. However, their DeepPRS percentile scores are 75, 98 and 86, which are larger than threshold of the 2.23-fold risk (**Supplementary Table S10**).

2.5. Each component of DeepPRS is well-designed for diseases risk prediction

In order to illustrate the contribution of each part of DeepPRS to the prediction performance, we test three varied methods by replacing three parts of the DeepPRS respectively, including using the traditional additive encoding method to replace our encoding schema (method a), directly using SNP features for subsequent feature extraction without the partial connected layer (method b), using CNN to replace BiLSTM to extract gene features (method c). In this test, we only use genotype features as input, and compare different methods based on ten-fold cross validation. The results

show that the performance of DeepPRS is better than the other three methods, indicating that the model is designed properly (**Table 3**). Compared with the additive encoding schema of method a, the reason for the superior performance of DeepPRS may be that the new encoding method expands the dimension of the feature. The traditional encoding method uses 0, 1 and 2 to represent genotypes, which is likely to bring additional quantitative relationships. In fact, there may not be a two-fold risk relationship between homozygous genotype and heterozygous genotype. We remove this additional quantitative relationship and better handle the missing genotype by filling it with (zero, zero) instead of deleting the locus. Compared with method b, DeepPRS uses partial connected layer to connect the SNP features to the neuron representing their adjacent gene. It not only greatly reduces the number of parameters and running time of deep neural networks model, but also takes the association between SNPs and genes into account. In addition, the effect of SNPs on the individuals is ultimately expressed by genes, and it is of more biological significance to extract features using gene-based partial connected layer. Compared with the CNN model of method c, the reason for the superior performance of DeepPRS may due to the fact that CNN can only extract local SNPs interaction relations, while BiLSTM can capture SNP interaction information at a distance on a chromosome.

Table 3. The effect of each component of DeepPRS on disease risk prediction.

	DeepPRS	Method a	Method b	Method c
Disease	AUC	AUC	AUC	AUC

Alzheimer's disease	0.7245	0.7158	0.7216	0.7232
Inflammatory bowel disease	0.6517	0.6485	0.6469	0.6473
Type 2 diabetes	0.6508	0.6490	0.6440	0.6457
Breast cancer	0.6227	0.6169	0.6232	0.6162

2.6. DeepPRS is robust for the variation of significant SNPs on disease risk prediction

At present, GWAS has not identified all significant loci. However, DeepPRS, pruning and thresholding method and Lasso method all need to select significant SNPs based on the P value given by the GWAS. Therefore, we test whether the absence of significant SNPs may affect the performance of these three methods. We sequentially removed the most significant SNPs in the top 10, top 20 and top 50 from GWAS result. The results show that DeepPRS performs consistently better than the other two methods (**Figure. 4**). Although the AUCs of all three methods decrease with the increase of missing significant SNPs, DeepPRS is the most robust one (**Figure. 4**). For example, the decrease of AUC for DeepPRS is more than two times slower than pruning and thresholding method on breast cancer (**Figure. 4d**). In summary, compared with the pruning and thresholding method, and lasso method, the results show that the DeepPRS is more robust for the variation of significant SNPs on disease risk prediction.

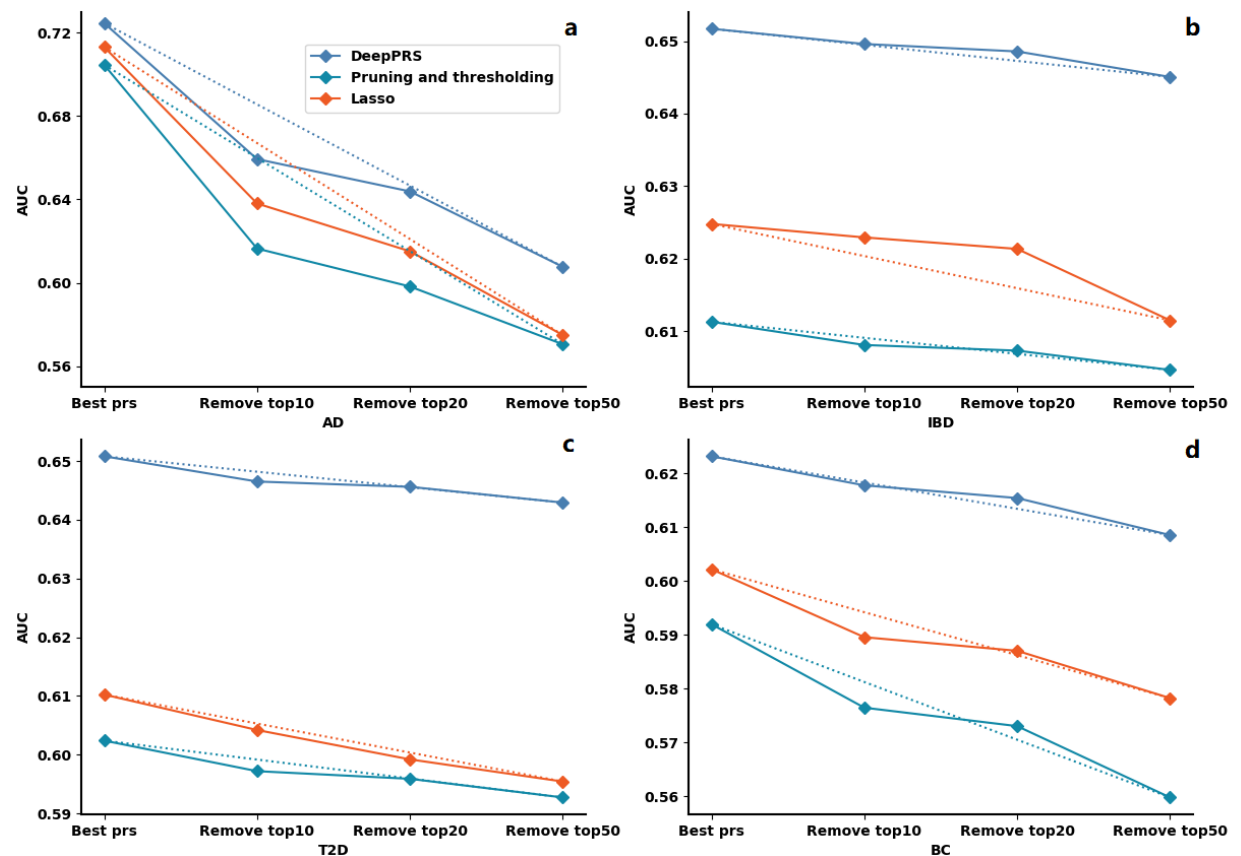


Figure 4. The results of risk prediction for Alzheimer's disease(a), inflammatory bowel disease (b), type 2 diabetes (c) and breast cancer (d) after removing different significant SNPs. The dotted line indicates the rate of performance degradation. The X axis represents the number of significant SNPs removed in different evaluation tests, and the Y axis represents the AUC score.

3. Discussion

In the past 25 years, the clinical potential of PRS has attracted widespread attention. The success of PRS depends on not only the genome-wide association study (GWAS) data but also the development of risk models. More accurate disease risk prediction may lead to better customize screening, prevention and treatment. In this study, we propose a novel PRS calculation model and analyze the risk of four common diseases. Alzheimer's disease, as the most common form of dementia, accounts for about 60% of

cases of dementia^[35]. At present, there are about 50 million AD patients worldwide, which has become one of the biggest public health challenges. Early diagnosis of AD may bring important personal and economic benefits^[36]. DeepPRS_{AD} identifies 7.90% of the population at greater than fivefold risk and the top 1% of the population have more than 8.97-fold risk (**Supplementary Table S1**). For AD, the proportion of individuals at high risk is much higher than the other three diseases, which may be due to the fact that the number of cases in the AD data set is too small. As a risk warning, this score can be a sign of early intervention for high-risk people. The results of a large, long-term, randomized controlled trial show that multifaceted interventions, including diet, exercise, cognitive training, and vascular risk monitoring, can improve or maintain cognitive function in the general dementia risk population^[37] (60-77 years old).

As a global disease with accelerating incidence in newly industrialized countries, the burden of IBD remains high and prevalence surpasses 0.3%^{[38][39]}. DeepPRS_{IBD} identifies 4.19% of the population at greater than threefold risk and the top 1% have more than 3.75-fold risk (**Supplementary Table S1**). Although existing studies have shown that factors such as diet, probiotics, and antibiotics are related to the development of IBD, more studies are still needed to explore the mechanisms that can help to prevent IBD^{[38][40]}. The high-risk population identify by DeepPRS may bring new opportunities for large-scale population epidemiological studies to assess a novel preventive therapy.

Type 2 diabetes is an expanding global health problem and puts a huge burden on health-care systems^[41]. DeepPRS_{T2D} identifies 4.10% of the population at greater than

threefold risk and the top 1% have more than 3.70-fold risk (**Supplementary Table S1**).

A study from Finland found that increasing physical activity and an intensive lifestyle may substantially reduce the incidence of type 2 diabetes in high-risk individuals^[42].

Therefore, the high-risk population determined by our DeepPRS_{T2D} can take early prevention to reduce the risk of type 2 diabetes.

Breast cancer is the second leading cause of death from cancer in women, affecting one in twenty globally and as many as one in eight in high-income countries^[43]. Fortunately, studies have shown that early detection and treatment can considerably improve outcomes of patients^{[43][44]}. DeepPRS_{BC} identifies 1.59% of the population at greater than threefold risk and the top 1% have more than 3.28-fold risk (**Supplementary Table S1**). Although the current prevention methods continue to increase, the overdiagnosis of screening, the serious side effects of chemical and biological prevention still cannot be ignored^[45]. Ascertainment of those with high deep polygenic risk score may provide an opportunity to adopt these interventions more precisely.

These results show that the DeepPRS can identify higher-risk individuals in the population for several common diseases. In addition to using the genotype information available at birth as a predictor, more clinical features can be added to the predictor over time. Then, the accuracy of the DeepPRS may be improved.

Acknowledgments

This research is conducted with approved access to UK Biobank data under application number 53464.

Funding:

National Natural Science Foundation of China grant No. 62072376 and U1811262

Author contributions:

Conceptualization: JJ.P, Z.W, X.S
Methodology: JJ.P, J.L, R.H, Z.W
Investigation: J.L, R.H, Y.W, L.H, JH.P
Supervision: JJ.P, T.W, J.H, X.S, Z.W
Writing—original draft: JJ.P, J.L, R.H, Y.W
Writing—review & editing: JJ.P, J.L, T.W, J.H, X.S, Z.W

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data used in the analyses are available in UK Biobank at <https://www.ukbiobank.ac.uk/>. The code will be publicly released after the paper is accepted.

References

- [1] Khera, A. V., Chaffin, M., Aragam, K. G., Haas, M. E., Roselli, C., Choi, S. H., ... & Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, 50(9), 1219-1224.
- [2] Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., ... & McCarthy, M. I. (2020). A brief history of human disease genetics. *Nature*, 577(7789), 179-189.

- [3] Carlson, C. S., Eberle, M. A., Kruglyak, L., & Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990), 446-452.
- [4] Gusella, James F., and Marcy E. MacDonald. "Huntington's disease and repeating trinucleotides." (1994): 1450-1451.
- [5] Kremer, B., Goldberg, P., Andrew, S. E., Theilmann, J., Telenius, H., Zeisler, J., ... & Hayden, M. R. (1994). A worldwide study of the Huntington's disease mutation: the sensitivity and specificity of measuring CAG repeats. *New England Journal of Medicine*, 330(20), 1401-1406.
- [6] Riordan, J. R., Rommens, J. M., Kerem, B. S., Alon, N., Rozmahel, R., Grzelczak, Z., ... & Chou, J. L. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, 245(4922), 1066-1073.
- [7] Kerem, B. S., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., ... & Tsui, L. C. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science*, 245(4922), 1073-1080.
- [8] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease[J]. *Nature genetics*, 2003, 33(3): 228-237.
- [9] Choi S W, Mak T S H, O'Reilly P F. Tutorial: a guide to performing polygenic risk score analyses[J]. *Nature Protocols*, 2020, 15(9): 2759-2772.

- [10] Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., ... & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720), 385-389.
- [11] Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., ... & McCarthy, M. I. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature genetics*, 50(11), 1505-1513.
- [12] Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., ... & Pietiläinen, K. H. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317), 832-838.
- [13] Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., ... & Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. *Nature communications*, 10(1), 1-9.
- [14] Dudbridge F. Power and predictive accuracy of polygenic risk scores[J]. *PLoS Genet*, 2013, 9(3): e1003348.
- [15] International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752 (2009).
- [16] Evans, D. M., Visscher, P. M. & Wray, N. R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.* 18, 3525–3531 (2009).

- [17]Torkamani, A., Wineinger, N.E. & Topol, E.J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* 19, 581–590 (2018).
<https://doi.org/10.1038/s41576-018-0018-x>.
- [18]Carlborg Ö, Haley C S. Epistasis: too often neglected in complex trait studies?[J]. *Nature Reviews Genetics*, 2004, 5(8): 618-625.
- [19]Hemani G, Knott S, Haley C. An evolutionary perspective on epistasis and the missing heritability[J]. *PLoS Genet*, 2013, 9(2): e1003295.
- [20]Strange, A., Capon, F., Spencer, C. C., Knight, J., Weale, M. E., Allen, M. H., ... & Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2. (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nature genetics*, 42(11), 985.
- [21]Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., & Park, J. H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature genetics*, 45(4), 400-405.
- [22]Lello, L., Raben, T. G., Yong, S. Y., Tellier, L. C., & Hsu, S. D. (2019). Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. *Scientific reports*, 9(1), 1-16.
- [23]Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211-252.

- [24] Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* 22, 1589–1604 (2017).
- [25] Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsdottir, B. J., Finucane, H. K., Salem, R. M., ... & Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3), 284.
- [26] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, 25(1), 24-29.
- [27] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203-209.
- [28] Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., ... & Nalls, M. A. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics*, 45(12), 1452-1458.
- [29] Liu, J. Z., Van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., ... & Weersma, R. K. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9), 979-986.

- [30] Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., ... & Gurrea, A. B. (2017). An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes*, 66(11), 2888-2902.
- [31] Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., ... & Ishiguro, J. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678), 92-94.
- [32] 1000 Genomes Project Consortium. A global reference for human genetic variation[J]. *Nature*, 2015, 526(7571): 68-74.
- [33] Bradley, Andrew P. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern recognition* 30.7 (1997): 1145-1159.
- [34] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection[C]//Ijcai. 1995, 14(2): 1137-1145.
- [35] Alzheimer's Association. 2019 Alzheimer's disease facts and figures[J]. *Alzheimer's & dementia*, 2019, 15(3): 321-387.
- [36] Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset[J]. *American journal of public health*, 1998, 88(9): 1337-1342.
- [37] Ngandu, T., Lehtisalo, J., Solomon, A., Levälahti, E., Ahtiluoto, S., Antikainen, R., ... & Kivipelto, M. (2015). A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. *The Lancet*, 385(9984), 2255-2263.

- [38]Ng, S. C., Shi, H. Y., Hamidi, N., Underwood, F. E., Tang, W., Benchimol, E. I., ... & Kaplan, G. G. (2017). Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *The Lancet*, 390(10114), 2769-2778.
- [39]Seyedian S S, Nokhostin F, Malamir M D. A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease[J]. *Journal of medicine and life*, 2019, 12(2): 113.
- [40]Jairath V, Feagan B G. Global burden of inflammatory bowel disease[J]. *The Lancet Gastroenterology & Hepatology*, 2020, 5(1): 2-3.
- [41]DeFronzo, R. A., Ferrannini, E., Groop, L., Henry, R. R., Herman, W. H., Holst, J. J., ... & Weiss, R. (2015). Type 2 diabetes mellitus. *Nature reviews Disease primers*, 1(1), 1-22.
- [42]Laaksonen, D. E., Lindström, J., Lakka, T. A., Eriksson, J. G., Niskanen, L., Wikström, K., ... & Uusitupa, M. (2005). Physical activity in the prevention of type 2 diabetes: the Finnish diabetes prevention study. *Diabetes*, 54(1), 158-165.
- [43]Britt, K.L., Cuzick, J. & Phillips, KA. Key steps for effective breast cancer prevention. *Nat Rev Cancer* 20, 417–436 (2020). <https://doi.org/10.1038/s41568-020-0266-x>
- [44]McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ... & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89-94.

[45]Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., ... & Zhu, H. P.

(2017). Risk factors and preventions of breast cancer. International journal of biological sciences, 13(11), 1387.