

Commentary

Research progress of polygenic genetic risk score for precise prevention

Wang Tianpei, Jin Guangfu, Hu Zhibin, Shen Hongbing 211166

Nanjing, Department of Epidemiology, School of Public Health, Nanjing Medical University (Wang Tianpei, Jin Guangfu, Hu Zhibin, Shen Hongbing);

215008 Suzhou, Suzhou College of Nanjing Medical University (Wang Tianpei, Jin Guangfu, Hu Zhibin) Corresponding author : Jin Guangfu, E-mail:

guangfujin@njmu.edu.cn DOI: 10.16462 / j.cnki.zhjbkk.2021.09.001

Abstract In recent years, genome-wide association studies have identified genetic susceptibility loci for a large number of complex diseases. Polygenic genetic risk scores have been shown to be useful for quantifying multiple complex diseases by integrating the effects of multiple susceptibility loci This paper introduces the construction and evaluation method of polygenic genetic risk score, and summarizes the latest research progress in the application of precision prevention. Key words: polygenic genetic risk score; genome-wide association study; precise prevention Article No. 16743679(2021)090993y05

Abstract In recent years, genome-wide association studies have identified genetic susceptibility loci for a large number of complex diseases. Polygenic genetic risk scores have been shown to be useful for quantifying multiple complex diseases by integrating the effects of multiple susceptibility loci This paper introduces the construction and evaluation method of polygenic genetic risk score, and summarizes the latest research progress in the application of precision prevention. Key words: polygenic genetic risk score; genome-wide association study; precise prevention Article No. 16743679(2021)090993y05

Abstract In recent years, genome-wide association studies have identified genetic susceptibility loci for a large number of complex diseases. Polygenic genetic risk scores have been shown to be useful for quantifying multiple complex diseases by integrating the effects of multiple susceptibility loci This paper introduces the construction and evaluation method of polygenic genetic risk score, and summarizes the latest research progress in the application of precision prevention. Key words: polygenic genetic risk score; genome-wide association study; precise prevention Article No. 16743679(2021)090993y05

According to statistics from the National Human Genome Research Institute of the United States and the European Institute of Bioinformatics, more than 15,000 genetic loci associated with diseases or traits have been identified by genome-wide association studies (GWAS) so far [1]. , established the link between genomic DNA sequence and human phenotype and disease, which is a major progress in the understanding of the genome after the completion of the Human Genome Project. Previous reviews have introduced various aspects of GWAS[2] , among which the identification of GWAS An important cognition of the genetic locus is that the effect of a single genetic locus on the occurrence of complex phenotypes/diseases is weak, that is, it reflects the characteristics of polygenic inheritance. Therefore, the effect of multiple genetic loci is integrated into polygenic genetic risk Polygenic risk score (PRS) is more representative of phenotype/disease

With the continuous accumulation of genomics data and the establishment of large-scale prospective population cohorts, such as the China Chronic Disease Cohort (China Kadoorie Biobank, CKB) [3] and the UK Biobank (UK Biobank, UKB) [4], etc., the predictive performance of PRS in a variety of complex diseases and phenotypes has been independently evaluated prospectively[5-7] It has been proved that it can be used as a genetic index to represent genetic strength or genetic risk for population risk stratification, which has potential application value in predicting disease risk, treatment selection, and disease prognosis estimation, and promoting the development of precision medicine. The construction and evaluation methods and their research progress in the application of precision prevention are briefly introduced to provide reference for related research.

1 PRS construction method

The classical construction method of PRS (clump and threshold, C + T method) combines the effects of risk alleles carried by individuals into a score that reflects individual disease susceptibility, including two important parts: the inclusion of genetic loci and the weighting. The researchers first screened the genetic loci and determined the weight of each locus according to the proposed criteria in GWAS, such as the P threshold of association analysis 5×10^{-6} , etc., such as using OR value for binary variable phenotypes. Natural logarithm (lnOR), using β value for continuous variable phenotypes, and combining the effects of the included genetic loci to obtain the individual PRS value. Earlier studies usually used stricter genome-wide association significance criteria.

Standard ($P < 5 \times 10^{-6}$), this criterion has the advantage of reducing the occurrence of Type I errors by a stringent criterion. However, the genetic loci included in this criterion explain only a small fraction of the heritability of the disease, thus limiting the predictive power of PRS. Several studies [5, 8] have shown that a looser P threshold criterion can improve the heritability explained by PRS and thus improve the predictive performance. At the same time, due to the linkage between single nucleotide polymorphisms (SNPs) on a genome-wide scale. Due to the existence of linkage disequilibrium (LD), researchers often use LD clumping or LD pruning to delete genetic loci in the same LD block to obtain independent genetic signals. Currently, software for such SNP screening methods is provided. The classical construction method of PRS selects some sites with the strongest correlation (the lowest P value) in GWAS to construct, which is prone to the phenomenon of "victor's curse", which leads to overfitting of the model and may reduce the external performance of PRS. Therefore, in recent years, a variety of methods have been devoted to including genome-wide loci, and the loci effects are adjusted by penalizing algorithms or incorporating LD information, locus function information and other methods to construct PRS. This research group in the early stage, five PRS calculation principles and methods were introduced [11]. In recent

years, the PRS construction methods have been further developed. The University of Michigan systematically compared and summarized the current 46 construction methods [12], which can be mainly divided into (1) based on the complete Bayesian method, using the Markov chain Monte Carlo method for model fitting, such as BSLMM [14] and other methods; (2) based on empirical Bayesian BayesR method, using such as LD relationship between sites, optimization of site function models such as LDpred information, AnnoPred [16], etc.; (3) based on frequentist methods, such as MultiBLUP, penalized regression, using adaptive algorithm for parameter estimation, such as lasso [20]. These methods have been Khera et al. [13] reduced the LD bias and CTR has been reported in 2014. The range based on the relationship between the largest sample size of obesity GWAS and the genetic locus LD of the 1000 Genomes Project European population. 2.1 million genetic loci to construct genome-wide PRS, then 100,000 in UKB

Validation was carried out in more than 300,000 individuals of European descent, and the predictive power of PRS on obesity was finally evaluated in 300,000 independent individuals. The results showed that the effect of PRS on individual obesity in adolescence increased with age, and at the age of 18 years PRS > P90. Compared with the other 90% of the population, the high-risk group gained 12 kg in weight, and the risk of severe obesity in the middle-aged high-risk group was 25 times that of the rest. Methods About 2.5 million genetic loci were included to construct pulmonary function PRS, and 10 case-control studies, cohort studies, and about 28 500 people were used to verify and evaluate the predictive power of PRS for chronic obstructive pulmonary disease. The results showed that the European PRS > P90 The risk of developing chronic obstructive pulmonary disease was 799 times higher in the high-risk group of PRS than in those in the lowest 10% of PRS, and 483 times in the non-European population.

2. Evaluation method of PRS prediction performance

At present, the commonly used PRS evaluation methods are as follows: (1) the goodness of fit of the model, the goodness of fit of the linear regression R^2 statistic (for continuous outcome variables) and the goodness of fit Nagelkerke in the logistic regression analysis model. The sR^2 statistic (for dichotomous outcome variables) is often used to evaluate the degree of interpretation of PRS for outcome variables. The higher the goodness of fit, the higher the degree of interpretation of PRS for outcome variables. The degree of discrimination is an indicator for judging the ability of a model to correctly distinguish people with different disease risks. The area under the curve (AUC) and the C index in the time-dependent receiver operating curve (ROC) analysis are often used. The results can be interpreted as the probability that randomly selected cases have higher PRS values than randomly selected controls, ranging from 0.5 to 1.0. The closer the AUC is to 1.0, the higher the probability of PRS prediction. Such models often include age, gender, and other known risk factors for disease. At the same time, net reclassification improvement (net reclassification improvement, NRI) and integrated discriminant improvement index (integrated discrimination improvement index, used to evaluate IDI) are also often used. The improvement of the model after additionally incorporating PRS into the model, if $NRI > 0$, the new model after incorporating PRS is a positive improvement over the old model, and vice versa. It indicates that the predictive ability of the new model is better. (2) Model calibration degree, which is an important indicator to evaluate the accuracy of the model in predicting the probability of a future outcome event for an individual. This method compares the consistency between the predicted risk and the observed risk by PRS. (3) The prediction ability of the model is to use the PRS score to divide the population into different disease risk groups. For example, the population is often divided into 10 or 5 equal parts according to the PRS, and the different risk groups are compared. The difference between the relative risk and the absolute risk of disease occurrence in a population.

3 Evaluation of the application of PRS in precise prevention

Identifying high-risk groups and promoting precise prevention is a key public health task. Traditional risk factors are often unmeasurable early in life and change significantly over time. The construction of PRS is based on germline genetic variation, which is present at birth. A genetic risk score can be obtained, so that the risk of developing the disease can be predicted early in an individual's life. Although an individual's genetic status is fixed, genetic risk is dynamic and depends on age, environmental exposure. Intervention or targeted prevention when individuals are not exposed to environmental risk factors and have not formed lifestyle habits can reduce the risk of individual disease occurrence, that is, individuals with high PRS are encouraged to take targeted prevention against the cause of the disease. Therefore, PRS can be used to guide the primary prevention of disease. In addition, the genetic stratification of populations indicated by PRS is important for disease screening and early diagnosis and treatment (ie, secondary prevention), as well as treatment decisions and predicting patient outcomes (ie, tertiary prevention).

3.1 Primary prevention

At present, a number of studies have established PRS to quantify the genetic risk of various clinically relevant traits and diseases. Mega et al[22] used 27 coronary heart disease patients identified in a previous large-scale GWAS study. The PRS was constructed from susceptibility loci and evaluated in a prospective cohort study, and the relative risk of coronary heart disease after taking statins was evaluated in a total of 48 421 individuals in 4 randomized controlled trials. The researchers found that PRS can quantify the risk of coronary heart disease, and the low genetic risk group (PRS<P20), the intermediate genetic risk group (PRS: P20 - P80), and the high genetic risk group (PRS>P80) were in The relative risk of coronary heart disease was reduced by 13%, 29% and 48% after statin use, respectively. The absolute risk of coronary heart disease in the low genetic risk group decreased from 30% to 19% after statin use, while the high genetic risk group From 66% to 36%, the number needed to treat (NNT) to prevent 1 case of coronary heart disease in the high genetic risk group was only 1/3 of that in the low genetic risk group. This evidence supports the inclusion of PRS The identification of individuals at high risk of cardiovascular disease can be optimized to maximize the benefits of prophylactic statin use.

This research group constructed and evaluated the application effect of gastric cancer PRS in risk prediction[8]: In the first stage, a multicenter large-sample GWAS study of gastric cancer was carried out in the Chinese population (including 10 254 gastric cancer cases and 10 914 cancer-free controls).), and according to the genetic association results ($P < 5 \times 10^{-8}$), 112 gastric cancer susceptibility loci in the Chinese population were screened out, and the polygenic risk score PRS₁₁₂ was constructed. The results showed that there was a dose-response relationship between PRS₁₁₂ and the risk of gastric cancer, and the high genetic risk group (PRS>P80) had a lower incidence of gastric cancer.

The risk of gastric cancer was 208 times higher than that of the risk group (PRS<P20); in addition, the risk of gastric cancer in the high genetic risk group who maintained a healthy lifestyle (not smoking, not drinking, eating less pickled vegetables, and eating more fresh fruits and vegetables) was higher than that of the high genetic risk group. 47% reduction in unhealthy lifestyles. Recently, our group further used the published genome-wide association study data to construct the PRS of 20 cancers, with the incidence of each cancer as the weight, and established the overall prevalence of cancer in men and women by gender.

Based on the cancer polygenic risk score (CPRS), 442,501 participants who applied UKB were evaluated. The study found that in high genetic risk populations (up to 20% in CPRS), maintaining a healthy lifestyle can reduce The 5-year absolute risk of cancer in men decreased from 7.23% to 5.51%, and from 5.77% to 3.69% in women. These results suggest that PRS should be used in primary prevention to screen high-benefit groups for targeted preventive interventions.

3.2 Secondary prevention of PRS in cancer patients

Secondary prevention of PRS in cancer patients also in cancer screening. According to the cancer screening developed by the US Preventive Medicine Working Group The screening guidelines recommend that women start breast cancer mammography screening at the age of 50, but considering individual risk factors, the screening initiation time can be advanced to 40-49 years of age [24] This age-based screening recommendation It was developed by comparing the average risk of breast cancer by age and balancing the risk of harm from false-positive results. Previous studies of breast cancer PRS found that approximately 16% of individuals in the population (high PRS with concomitant traditional risk factors)40 The risk of breast cancer at the age of 50

the average risk of the general population at the age of 50, which can prompt them to participate in breast cancer screening in advance. At the same time, individuals with low PRS and no traditional risk factors account for about 32% of the general population, and this part of the population is 50 years old. The risk of breast cancer at 40 years of age was lower than the average risk of the general population at age 40, suggesting that they do not need to participate in screening in advance. These results suggest that PRS can help

determine the starting age of targeted screening for different genetic risk groups. et al[25]

further evaluated the risk-benefit ratio and cost-effectiveness of PRS risk stratification in screening based on a breast cancer screening cohort study in the United Kingdom. The study found that with the expansion of the screening population, the population quality-adjusted life years (The increase in quality, adjusted life, years, QALYs) tends to be flat, while the cost of screening and overdiagnosis increase. When screening is limited to high-risk women with PRS > P70, breast cancer screening per 10,000 visits compared Age-based screening programs reduce costs by approximately \$720,000, overdiagnosis by 714%, and breast cancer deaths by 96%, while QALYs increase by approximately 44%. 3.

Forgetta et al. [26] quantified the genetic risk of fracture by constructing a PRS, and confirmed that the inclusion of PRS can greatly reduce the number of bone mineral density examinations under the premise of maintaining the sensitivity at 934% and the specificity at 985%, thereby improving bone quality. The efficiency of schizophrenia screening. Vassos et al [27] constructed a PRS for schizophrenia, and proved it through research.

In fact, PRS helps to differentiate schizophrenia patients from other psychiatric diagnoses in first-episode psychosis patients, and high-risk patients with PRS>P80 double the risk of being diagnosed with schizophrenia, suggesting that PRS can be used for psychiatric diagnosis. Auxiliary diagnosis for patients with schizophrenia. 3.3 Tertiary prevention pharmacogenetic studies test how genetic variation affects the response to treatment to assist in the selection of treatment methods to maximize efficacy and reduce side effects. In 2018, Ward et al. [28] constructed The PRS for depression was evaluated and evaluated in three treatment cohorts. The study showed that the population with higher PRS had a better trend of antidepressant treatment with a smaller sample size. A study[29] constructed the PRS for schizophrenia and further analyzed the two-way PRS. To evaluate whether PRS is related to the treatment response to the first-line mood stabilizing drug—lithium in patients with affective disorders. The results of the study showed that the lower the PRS score, the stronger the patient's response to lithium treatment. Compared with the high-risk group, the OR was 346. The study of Rush et al. [30] showed that only about one-third of the depression patients responded to the first antidepressant prescribed by the doctor, suggesting that the use of PRS and other antidepressants Risk factors are of great significance in guiding treatment selection. The papers in the "Tumor Epidemiology" column of this issue are the results of our team's research on the identification of tumor susceptibility loci and the construction and application of polygenic risk scores. Li Qian et al. [31] A new gastric cancer susceptibility genetic variant rs2517714 was identified through fine mapping analysis of the major histocompatibility complex (MHC) region in the Chinese Han population, and it was found through systematic functional annotation that it may affect the susceptibility gene HLA. This result will provide new evidence for the determination of genetic markers in the MHC region and the optimization of gastric cancer PRS. Hu Beiping [32] and Zhu Mengyi et al[33] used the previously constructed tumor polygenes The risk score CPRS was used in the UKB database to explore the interaction of blood lipid levels, C-reactive protein (CRP) and genetic risk in tumor pathogenesis. In males with different genetic risks, abnormal blood lipid levels all increase the risk of cancer in males. Zhu Mengyi et al. [33] found that elevated CRP mainly increases the risk of cancer in low- and medium-genetic risk groups, and CRP and genetic risk are significantly higher in males. There is a negative multiplicative interaction in tumor incidence. These two studies used the constructed CPRS as a measurable indicator to represent genetic factors for tumor epidemiology and etiology research, providing a reference for similar studies. As mentioned above, with the increase of GWAS sample size and the emergence of large cohort population data, PRS research has However, the effective use of PRS requires appropriate construction and evaluation methods, and reasonable interpretation of the research results. For this reason, Wand et al. [34] proposed standards for PRS-related research. , PRS is highly sensitive to the ethnic background of the population, and different populations

There may be differences in the allele frequency, LD relationship and locus effect of genetic loci between different ethnic groups, resulting in a decrease in the predictive power of PRS in different ethnic groups. Currently, the population and genetic loci of GWAS studies are mainly derived from European and American populations, including Chinese The East Asian population, including the population, only accounts for about discussed its possible clinical impact and benefits, and the distance of PRS from clinical application still needs to be evaluated by population trials and health economics methods. The tens of thousands of genetic loci discovered by GWAS reveal the relationship between genome and disease, which is a major contribution after the completion of the Human Genome Project. PRS is an important breakthrough for the application of precision medicine. Major progress will lay the foundation for precise prevention of complex diseases, including chronic diseases. There is no conflict of interest.

references

[1] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

[2] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

[3] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

[4] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

[5] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

[6] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

[7] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

[8] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

[9] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

[10] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

[11] Wang Cheng, Dai Juncheng, Sun Yimin, et al. Principles and methods of genetic risk scoring [J]. Chinese Journal of Epidemiology, 2015, 36(10): 1062 – 1064. DOI: 10.3760/j.issn.0512-2465.2015.10.1062.

