Accurate Prediction Model - Polygenic Risk Score for High-Risk Individuals Predictive of Gastric Cancer

Fujiao Duan
Zhengzhou University
Chunhua Song
Zhengzhou University
Peng Wang
Zhengzhou University
Hua Ye
Zhengzhou University
Liping Dai
Zhengzhou University
Jianying Zhang
Zhengzhou University
Kaijuan Wang (wkj@zzu.edu.cn)
Zhengzhou University https://orcid.org/0000-0002-3300-9453

Research Article

Keywords: Gastric cancer, Incidence risk, Risk factor, Polygenic risk score, Risk prediction model

Posted Date: November 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1048838/v1

Abstract

Background

The genetic variation of gastric cancer has not been fully identified. We aimed to screen and identify common variant single nucleotide polymorphisms (SNPs) and long noncoding RNA (lncRNA) related SNPs associated with the risk of gastric cancer, and construct and evaluate prediction models based on polygenic risk score (PRS).

Methods

Non-genetic factors such as *H.pylori* infection, environment, and genetic factors associated with gastric cancer were screened following meta-analysis and bioinformatics,verified by frequency matched case-control study. PRS and weighted genetic risk scores (wGRS) were derived from estimation of effect size. Net reclassification improvement (NRI), integrated discrimination improvement (IDI), akaike information criterion (AIC) and bayesian information criterion (BIC) were used to evaluate model.

Results

A risk gradient was observed across quantile of the PRS, the results showed that the risk of gastric cancer in the highest 10 quantile of PRS was 3.24 folds higher than that of the general population (*OR*=3.24,95%CI: 2.07, 5.06). The PRS with one or more risk factors (smoking, drinking and *H. pylori* infection) was superior to the single genetic risk model. For NRI and IDI, the PRS combinations were significantly improved compared to wGRS model combinations (*P*<0.001). The model of PRS combined with lncRNA SNPs, smoking, drinking and *H. pylori* infection was the best fitting model (AIC=117.23, BIC=122.31).

Conclusion

Our findings indicated that the model based on PRS combined with lncRNA SNPs, smoking, drinking, and *H. pylori* infection had the optimal predictive ability on the risk of gastric cancer, contributing to distinguish high-risk groups from population.

Background

Gastric cancer is a highly lethal cancer worldwide, being the fourth most common malignancy and the third leading cause of cancer-related mortality in developing countries[1, 2]. According to the cancer statistics released by cancer registration center in China, gastric cancer is the second most commonly diagnosed cancer among men, second only to lung cancer[3].

Studies have shown that the occurrence and development of gastric cancer is a multifactor and multistage process, which is highly related to biological, environmental exposure and genetic factors. *Helicobacter pylori* (*H. pylori*) is a well-known biological pathogenic factor[4]. The average infection rate of *H. pylori* in Chinese population is as high as 60%[5], which is also one of the causes of high incidence rate of gastric cancer in China.

The polygenic risk score (PRS) is a risk prediction models based on demographic and clinical characteristics, genetic markers and other risk hierarchical factors. At present, studies have shown that PRS has promising predictive power for cancer high-risk group[6]. Provide clinicians with a tool that enables them to assess the risk of patients and improve the utilization of medical resources[7, 8].

The PRS model incorporating genetic and non-genetic factors has robust risk prediction capability, showing that there is an interaction (multiplier effect) between breast cancer-related single nucleotide polymorphisms (SNPs) and environmental risk factors[9]. Other similar studies have also confirmed the remarkable predictive power of PRS[10–12] This method has also been utilized to predict the risk of psoriasis[13], stroke[14] schizophrenia, and bipolar disorder[15, 16]. The construction of the model has achieved good prediction results.

The genome-wide association studies (GWASs) have shown that there are about 11 loci in the human genome associated with gastric cancer[17]. The previously developed weighted genetic risk scores (wGRS) has some limitations in predicting cancer risk, and models based on wGRS were basically depend on genetic sites screened by GWAS or evidence-based medicine (EBM) [18–20]. So far, the construction of risk prediction model based on PRS for gastric cancer research has not been reported in China. In other complex disease studies, it has been shown that the prediction ability of risk prediction model based on PRS was better than that based on wGRS[7, 8, 14]. Meanwhile, studies have confirmed that lncRNA SNPs were associated with gastric cancer. However, lncRNA SNPs were not found in the existing risk prediction model[21–23].

In the present study, quantitative systematic evaluation and meta-analysis were used to determine the non-genetic and genetic factors such as *H. pylori* infection and environment. According to the results of association and bioinformatics, gastric cancer-related SNPs were screened and verified by case-control study. Based on the validation results, lncRNA SNPs, as an independent risk factor data set, combined the common SNPs with *H. pylori* infection and environmental factors by PRS to construct an individualized risk prediction model for gastric cancer.

Materials And Methods

The study was approved by the ethics committee of Zhengzhou University. All participants were informed and signed written informed consent. The design and implementation flow chart of this study was shown in Figure 1.

Meta-analysis of risk factors for gastric cancer

To obtain the credibility and strength of non-genetic factors and genetic variation on gastric cancer risk, we performed a field synopsis and meta-analysis to identify the risk of gastric cancer in Chinese population. A total of 22 SNPs involving 16 genes were identified and associated with the risk of gastric cancer. Details have been published in the journal of Aging-US[24]

Genetic variant selection for PRS

The bioinformatics method was used to screen lncRNAs and corresponding functional SNPs that were differentially expressed in gastric cancer and possess potential binding sites with microRNAs (miRNAs).

The gastric cancer related microarray data (gse50710, gse53137, gse58828) of Chinese population in the Gene Expression Omnibus (GEO) database were retrieved and downloaded. The GEO chip data related to gastric cancer was analyzed by using the Bioconductor software based on R-software (version 3.6.2 for Windows), which was associated to the mapping database of chip probes according to the probe code. The intersection part was obtained according to the analysis results of three chips by using SAS 9.2 (SAS Institute Inc., Cary, North Carolina, USA). The difference multiple was > 2.0 and $P < 0.05$, the differentially expressed lncRNAs were screened.

We used the lncRNASNP2 database (http://bioinfo.life.hust.edu.cn/lncRNASNP#!/) and the online database RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi), the preliminary potential function prediction of the biological functions of the SNPs on the differentially expressed lncRNAs were screened out, and the SNPs that affect the secondary structure of lncRNAs or affect the binding of miRNAs will be identified and screened out. r^2 can reflect the degree of linkage disequilibrium (LD) between SNPs sites, combined with the LD (r^2<0.8 and LD<1.0) between SNP sites on the same gene, 21 lncRNA SNPs were finally selected (supplementary Table 1).

We followed the principle of evidence-based medicine and applied a three-step approach. We initially performed meta-analysis to screen the genetic associations between genetic variant and gastric cancer. After this screening analysis, SNPs in strong linkage disequilibrium (LD) with each other polymorphisms ware excluded. Finally, the extracted SNPs were combined with the published field synopsis or systematic review on SNPs ($OR \geq 1.20$ or $OR \leq 0.8$) significantly associated with gastric cancer in Chinese population (Chinese Han in Beijing, Minor Allele Frequency≥0.1). Finally, a total of 18 genes involved in 20 SNPs were selected, the results were presented in supplementary Table 2.

Study population

All patients with gastric cancer were new cases from the First Affiliated Hospital of Zhengzhou University and the Affiliated Cancer Hospital of Zhengzhou University from January 2012 to December 2015. The patients did not receive anti-tumor treatment before recruitment, and had no history of other malignant tumors.

The controls were collected from a cardiovascular disease epidemiological survey conducted at the same time in Henan Province. Individuals with malignant tumors, digestive system diseases, and blood related to the case were excluded.

Based on frequency matched case-control study design to match subjects according to gender and age (± 2 years), the blood samples of 660 patients with gastric cancer confirmed by pathology and 660 normal controls from community were collected. Each participant met the requirements of the institutional review committee and gave informed consent.

Genotyping and quality control

Polymerase chain reaction restriction fragment length polymorphism (PCR-RFLP), created restriction site-PCR-RFLP (CRS-PCR-RFLP) and Improved Multiplex Ligation Detection Reaction (iMLDR^{TM}) were used to genotype SNPs corresponding to lncRNAs or selected by EBM. For iMLDR^{TM}, 3130XL sequencer (AppliedBiosystems, USA) was used for sequencing, and the GeneMapper 4.0 was applied to identify genotypes.

For PCR-RFLP typing, 10% of the samples were randomly selected and the sequencing results were compared with the experimental results. When the agarose gel electrophoresis pattern could not accurately determine the genotype, repeated experiments or direct sequencing were used to determine the genotype.

In the iMLDR typing test, agarose gel electrophoresis was used to detect each sample before typing, and 3% double blind sample quality control and negative control quality control. For quality control samples, the success rate (call rate) and accuracy rate were ensured to be more than 98%.

Weighted genetic risk scores

The population average risk (Genetic score) of single SNP was calculated based on the genotype frequency of the genetic variation and the *OR* of the meta-analysis in the Chinese population.

Genetic score (W) = $(1-p)^2+2p(1-p)OR+p^2OR^2$

p was the risk allele frequency.

Assuming that the genotypes of a SNP are AA, AB and BB, B is the risk allele, A is the non-risk allele, and the corresponding risk values are 1, *OR* and OR^2, then the weighted genetic risk scores (wGRS) is estimated as follows:

AA=1/W，AB=OR/W，BB=OR^2/W

wGRS=SNP1×SNP2×SNP3…… SNPn (Missing value set to 1)

Polygenic risk score

We derived a PRS specific to Chinese populations from all SNPs that have been verified to be associated with gastric cancer risk at genome-wide significance level. The PRS was constructed for cases and controls by summing the risk allele counts (i.e., subjects have 0, 1, or 2 risk alleles) for the associated variants weighted by their natural log transformed (ie, the ln of the odds ratios (*OR*)) effect sizes (*OR*) extracted from results of multivariate logistic regression model. For each participant, we summed the weighted risk allele counts and then divided the total number of loci to derive a mean weighted score, and the mean weighted score as the reference.

$$PRS_j = \sum_i^j n_{ij} \ln(OR_i)$$

j is the number of SNPs included in the model; nij is the number of the i-th risk allele (0, 1 or 2); *OR*i is the associated risk value *(OR*) between the risk allele of the i-th SNP and gastric cancer.

Statistical analysis

The Hardy-Weinberg equilibrium (HWE) test was performed on the genotype distribution of the control using Chi square test of goodness of fit. Unconditional logistic regression was used to implement the correlation analysis between the targeted SNPs and gastric cancer risk.

Plink 1.9 (NIH-NIDDK's Laboratory of Biological Modeling, Harvard University) was used for quality control of related SNPs, association analysis of allele and generation of PRSice-2 (Gavin Band, New York, USA) basic dataset and target dataset. Gastric cancer risk prediction models were constructed using SNP screened by EBM and verified by association based on wGRS and PRS. lncRNAs SNPs were put into the prediction models as independent datasets of risk factors and empirical *P*-value was used to perform 10,000 fittings within the model to optimize model parameters and build the optimal model.

Receiver operating characteristic (ROC) and area under curve (AUC) were used to evaluate the gastric cancer recognition degree of different models. Net reclassification improvement (NRI) and integrated discrimination improvement (IDI) were used to evaluate the predictive ability of wGRS and PRS models, akaike information criterion (AIC) and bayesian information criterion (BIC) were used to evaluate the fitting degree of the model.

Statistical analysis was performed with R software (version 3.6.1; The R Foundation for Statistical Computing, Vienna, Austria) and Stata version 13.1MP (StataCorp: College Station, TX, USA). A *p*-value of <0.05 was considered statistically significant with two-sided.

Results

Characteristics of subjects

For confirmation of gastric cancer risk factors, 660 gastric cancer patients and 660 healthy subjects were recruited in case-control studies. Age and sex were well balanced in individuals between the cases and controls. Cases had a more common family history of gastric cancer (11.93% versus 2.12%, *P*<0.001), and a higher proportion of ever-smokers (65.7% versus 53.0%, *P*<0.001), drinkers (30.61% versus 23.18%, *P*=0.005), and *H.*pylori infection (62.92% versus 49.46%, *P*=0.005) compared with controls (Table 1).

Screening and identification of SNPs related to gastric cancer

Multivariate non-conditional logistic regression analysis was used to explore the association between lncRNA SNPs and the risk of gastric cancer based on five genetic models (allele, heterozygous, homozygous, dominant and recessive), adjusted by gender, age, smoking, drinking and family history. The results show that 14 SNPs were found to be related to risk of gastric cancer in 21 associated lncRNA SNPs (Table 2). Stepwise logistic regression analysis was used to institute PRS related to SNPs and 15 SNPs were found among 20 genotyping SNPs (*P*<0.05)(Table 3).

Distribution of genetic risk score

For each lncRNA SNPs and common SNPs, PRS was approximately normal distribution (Supplementary figure 1). In lncRNA SNPs and common SNPs, wGRS and PRS in the case group were significantly higher than those in the control group (Supplementary figure 2 and figure 4).

Construction of PRS risk prediction model

For wGRS, the risk of gastric cancer was significantly elevated with the increase of score groups. According to wGRS distribution, the risk of gastric cancer increased significantly with the increase of score, taking 0-1 group as reference (Supplementary Table 3).

PRS measured by bar plot showed the variance ratio of the correlation results obtained under different *p*-value thresholds (P_t), that was, the distribution of the explanatory value (R^2) of the estimated phenotypic variation. Figure 2A showed the R^2 value (vertical axis) of the phenotypic variation of the PRS model under different P_t values (horizontal axis), and the highest point in the column graph indicated the optimal model. When P_t = 0.0818, the model was optimal, and about 8.8% of the cases were from genetic variation ($P = 6.4 \times 10^{-19}$).

The output results of PRSice-2 showed the fitted *p*-value distribution corresponding to the correlation results obtained under different P_t with the results of high-resolution plot. In this model, the best P_t value was the point with the highest line (the point with the minimum fitting *p*-value), and the P_t was 0.0818.

According to the distribution of the data set, the visual output data was divided into 10 groups, with 40-60 quantile as a reference. We calculated the PRS for each subject and then divided the subjects into nine groups. The increase of PRS was associated with a significantly increased risk of gastric cancer (Figure 3). A risk gradient was observed across quantile of the PRS. The results showed that the risk of gastric cancer in the lowest 10 quantile of the risk score was 47.1% lower than that of the general population (*OR*=0.53, *95%CI*, 0.34, 0.87), and the risk of gastric cancer in the highest 10 quantile of PRS was 3.24 folds higher than that of the general population (*95% CI*: 2.07, 5.06) (Figure 3).

Goodness of fit and model evaluation of risk prediction model

The NRI and IDI were used to estimate the improvement of the prediction effect of the model and compared the NRI and IDI values of four different models after adding one or more new risk factors. According to NRI, the results showed that the increase of gastric cancer risk was not statistically significant (positive improvement was 4%) except wGRS vs. PRS + lncRNA SNPs in the comparison of four different models. Based on IDI, the prediction effect of PRS combined model was significantly higher than wGRS combined model (P<0.001), and the increase of NRI was statistically significant in the other three models (Table 4).

By introducing different factors to compare AIC and BIC based on wGRS and PRS, the best fitting degree was selected. wGRS and PRS of simple genetic model as reference, the results showed that the model coupled with lncRNA SNPs was better than the single genetic models of wGRS and PRS. The PRS with one or more risk factors (smoking, drinking and *H.* pylori infection) was superior to the single genetic risk model. The model of PRS combined with lncRNA SNPs, smoking, drinking and *H.* pylori infection was the best fitting model (AIC = 117.23, BIC = 122.31) (Table 5).

According to the ROC curve, AUC results showed that the introduction of lncRNA SNPs on the basis of wGRS and PRS could significantly improve the prediction ability of the model (Figure 4). On the basis of simple genetic model, the introduction of gastric cancer related lncRNA SNPs, drinking alcohol and *H.*pylori infection could significantly improve the prediction ability of the model (Table 5). Based on the above evaluation index results, the model including lncRNA SNPs, smoking, drinking, *H. pylori* infection on the basis of PRS has the best predictive ability. This model combination has the same results as the best model goodness of fit combination.

Discussion

Due to the high heterogeneity of cancer phenotypes and the complexity of their etiology, individuals exposed to the same environment may have different risks[25].

The heterogeneity was the result of the interaction of different mechanisms and multiple conditions. Genetic risk models could quantitatively explain the heritability of phenotypes to a certain extent. In practical application, the index effect of a single genetic phenotype prediction needs to be combined into a comprehensive index to facilitate the screening of high-risk population or screening application.

On the basis of comprehensive quantitative evaluation of biological, environmental, behavioral and genetic factors and gastric cancer risk, combined with systematic review and meta-analysis published in authoritative journals related to gastric cancer in Chinese population. This study verified the selected 20 SNPs related to gastric cancer through a population-based case-control study, and constructed a risk prediction model according to the results of association verification. Meanwhile, we used bioinformatics method to screen lncRNAs related to gastric cancer, and determined the functional SNPs and verified them in the population. As an independent risk factor, it was included in the risk prediction model combined with environmental, behavioral and biological factors, and the model was evaluated in multiple dimensions according to the different risk factors included.

So far, wGRS has been widely used for weighting the genetic association effect value of SNPs[26]. As the occurrence of cancer is regulated by multiple genes and multiple loci, the genetic efficiency of few genes or even single gene is weak, so it is impossible to accurately predict the disease. Therefore, it is necessary to integrate the genetic information of multiple loci and polygenes. PRS is a new strategy to study the genetic susceptibility of cancer or other complex diseases. PRS can be used to quantify the cumulative effect of multiple loci or genes, and redefine the risk scale to accurately measure the disease susceptibility score of an individual[27], it considers the size of individual SNPs effect. At the same time, some studies have confirmed that when considering the combination of multiple loci, the pattern recognition is the highest[28–30]. Therefore, wGRS and PRS methods were used to construct the model, and comparative evaluation and analysis were carried out in this study.

The recent large-scale GWASs of gastric cancer have provided an opportunity to develop risk models that include genetic variation. Studies have been reported that PRS using the extended set of genetic variants can improve discrimination breast cancer[31], coronary artery disease[32], prostate cancer[33] and esophageal adenocarcinoma [34]. However, there is no empirical evidence for the effectiveness of genetic risk stratification in gastric cancer. In this study, PRS was associated with risks of gastric cancer, accounting for 8.18% phenotypic variance of gastric cancer. Those in the highest quantile of the PRS had 224% increased risk for gastric cancer.

In order to evaluate the risk prediction model constructed by wGRS and PRS, NRI and IDI were used to estimate the improvement of the prediction effect after adding one or more new risk factors. Under the same factors and conditions, through the calculation and comparison of NRI and IDI values of four different combinations of models, the results showed that each model combination of PRS was better than that of corresponding wGRS. The findings further verified the previous results and were consistent with the results of previous studies[8, 14]. Meanwhile, the results showed that gastric cancer related lncRNA SNPs could effectively improve the recognition of the model.

AIC and BIC are used to measure the goodness of statistical model fitting. Both of them introduce the penalty term related to the number of model parameters. The penalty term of BIC is larger than that of AIC. Considering the number of samples, when the number of samples is too large, it can effectively prevent the model complexity caused by the high precision of the model. In order to measure the goodness of fit, this study compared AIC and BIC based on wGRS and PRS by introducing different factors, and screened out the optimal one. wGRS and PRS genetic model as reference, the results showed that the model with lncRNA was better than that of wGRS and PRS alone, and the single or combined models of smoking, drinking and *H.pylori* infection were better than the single genetic risk model. The model of PRS combined with lncRNA SNPs, smoking, drinking and *H. pylori* infection was the optimal model (AIC = 117.228, BIC = 122.314). According to the included risk factors, single or combined introduction of gastric cancer related lncRNA SNPs, drinking and *H.pylori* infection on the basis of genetic model could significantly improve the predictive ability of the model.

According to the ROC curve and AUC results of the PRS, single or combined introduction of gastric cancer related lncRNA SNPs, drinking and *H.pylori* infection on the basis of simple genetic model could significantly improve the prediction ability of PRS. This cumulative effect of PRS with environmental and/or biological factors has been confirmed in other cancers[29, 30, 35].

In the present study, the risk identification provided by genetic profile was summarized in PRS and further improved by combining biological, behavioral and environmental factors. However, some limitations of this study should be pointed out. First, in the SNPs screening stage, the pooled results showed that there might be heterogeneity among the included studies. However, we did not explore the source of heterogeneity, which was close to the effect value and population validation, suggesting that the results should be broadly applicable. Second, we only evaluated the 20 risk loci identified in a Chinese population, which limits our findings from being generalizable to other ethnic populations with different effect sizes of variants, LD patterns, and allele frequencies. Third, the interactions among the independent risk factors included in PRS, such as the interaction between environmental and genetic factors, were not dealt with, and the interaction between SNPs was ignored in the analysis of multiple genes. Finally, only genome-wide significant mutations were selected to produce PRS, other loci have not been identified, rare and copy number variations with greater correlation should be included in future PRS.

Conclusion

In summary, gastric cancer related common SNPs and lncRNA SNPs had a significant combined effect. Under the same factors and conditions, the PRS model was better than the wGRS model and the introduction of gastric cancer related lncRNA SNPs could significantly improve the recognition. The model based on PRS combined with lncRNA SNPs, smoking, drinking, and *H. pylori* infection had the best predictive ability on risk of gastric cancer, contributing to distinguish high-risk groups from general population. This study has important practical significance for China to formulate accurate and efficient screening strategies, and improve the detection rate of early gastric cancer.

Abbreviations

lncRNAs: long non-coding RNAs; SNPs: Single nucleotide polymorphisms; PRS: Polygenic risk scores; wGRS: weighted genetic risk scores; NRI: Net reclassification improvement; IDI: Integrated discrimination improvement; AIC: Akaike information criterion; BIC: Bayesian information criterion; OR: Odds ratio; *H. pylori*: *Helicobacter pylori*; GWAS: Genome-wide association studies; GRS: Genetic risk scores; EBM: Evidence-based medicine; PCR-RFLP: Polymerase chain reaction restriction fragment length polymorphism; CRS-PCR-RFLP: Created restriction site-PCR-RFLP; HWE: Hardy-Weinberg equilibrium; ROC: Receiver operating characteristic; AUC: Area under curve; CHD: Coronary heart disease; RCT: Randomized controlled trials; LD: Linkage disequilibrium.

Declarations

Author Contributions

All authors contributed significantly to this work. FJD and KJW designed and drafted the manuscript. JYZ and LPD collected studies and summarized data. HY, CHS and PW copyedited manuscript, did statistical work and prepared figures. All authors reviewed this manuscript and approved the final draft.

Acknowledgements

Not applicable.

Competing interests

The authors declare no competing interests.

Availability of data and materials

All data generated or analysed during this study are included in this published article.

Consent for publication

None of the individual person's data is in this text for publication.

Ethics approval and consent to participate

This study was approved by the ethics committee of Zhengzhou University., and performed in accordance with the Declaration of Helsinki.

Funding

This work was funded by the National Natural Science Foundation of China (81373097), the National Science and Technology Major Project of China (2018ZX10302205), and the Major Project of Science and Technology in Henan Province (161100311400).

References

1. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu XQ, He J: **Cancer statistics in China, 2015**. *CA: a cancer journal for clinicians* 2016, **66**(2):115–132.
2. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ: **Cancer statistics, 2009 CA**. *Ca A Cancer Journal for Clinicians* 2003, **53**(1):5–26.
3. Chen W, Zheng R, Zhang S, Zeng H, Zuo T, Xia C, Yang Z, He J: **Cancer incidence and mortality in China in 2013: an analysis based on urbanization level**. *Chinese journal of cancer research = Chung-kuo yen cheng yen chiu* 2017, **29**(1):1–10.
4. Plummer M, Franceschi S, Vignat J, Forman D, de Martel C: **Global burden of gastric cancer attributable to Helicobacter pylori**. *International journal of cancer Journal international du cancer* 2015, **136**(2):487–490.
5. Liu J, Wang Y, Zhao Q, Luo R, Xiao M, Zhang M, Xie W: **Prevalence and risk factors for Helicobacter pylori infection in southwest China: a study of health examination participants based on 13C-urea breath test**. *Turkish journal of medical sciences* 2017, **47**(5):1456–1462.
6. Torkamani A, Wineinger NE, Topol EJ: **The personal and clinical utility of polygenic risk scores**. *Nature reviews Genetics* 2018, **19**(9):581–590.
7. Hachiya T, Kamatani Y, Takahashi A, Hata J, Furukawa R, Shiwa Y, Yamaji T, Hara M, Tanno K, Ohmomo H *et al*: **Genetic Predisposition to Ischemic Stroke: A Polygenic Risk Score**. *Stroke* 2017, **48**(2):253–258.
8. Chatterjee N, Shi J, Garcia-Closas M: **Developing and evaluating polygenic risk prediction models for stratified disease prevention**. *Nature reviews Genetics* 2016, **17**(7):392–406.
9. Rudolph A, Song M, Brook MN, Milne RL, Mavaddat N, Michailidou K, Bolla MK, Wang Q, Dennis J, Wilcox AN *et al*: **Joint associations of a polygenic risk score and environmental risk factors for breast cancer in the Breast Cancer Association Consortium**. *International journal of epidemiology* 2018, **47**(2):526–536.
10. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, Schumacher FR, Anderson WF, Check D, Chattopadhyay S *et al*: **Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States**. *JAMA oncology* 2016, **2**(10):1295–1302.
11. Shieh Y, Hu D, Ma L, Huntsman S, Gard CC, Leung JW, Tice JA, Vachon CM, Cummings SR, Kerlikowske K *et al*: **Breast cancer risk prediction using a clinical risk model and polygenic risk score**. *Breast cancer research and treatment* 2016, **159**(3):513–525.
12. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, Babb de Villiers C, Izquierdo A, Simard J, Schmidt MK *et al*: **BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors**. *Genetics in medicine: official journal of the American College of Medical Genetics* 2019, **21**(8):1708–1718.
13. Yin X, Cheng H, Lin Y, Wineinger NE, Zhou F, Sheng Y, Yang C, Li P, Li F, Shen C *et al*: **A weighted polygenic risk score using 14 known susceptibility variants to estimate risk and age onset of psoriasis in Han Chinese**. *PloS one* 2015, **10**(5):e0125369.
14. Reginsson GW, Ingason A, Euesden J, Bjornsdottir G, Olafsson S, Sigurdsson E, Oskarsson H, Tyrfingsson T, Runarsdottir V, Hansdottir I *et al*: **Polygenic risk scores for schizophrenia and bipolar disorder associate with addiction**. *Addiction biology* 2018, **23**(1):485–492.
15. Power RA, Steinberg S, Bjornsdottir G, Rietveld CA, Abdellaoui A, Nivard MM, Johannesson M, Galesloot TE, Hottenga JJ, Willemsen G *et al*: **Polygenic risk scores for schizophrenia and bipolar disorder predict creativity**. *Nature neuroscience* 2015, **18**(7):953–955.
16. Dima D, Breen G: **Polygenic risk scores in imaging genetics: Usefulness and applications**. *J Psychopharmacol* 2015, **29**(8):867–871.
17. Tian J, Liu G, Zuo C, Liu C, He W, Chen H: **Genetic polymorphisms and gastric cancer risk: a comprehensive review synopsis from meta-analysis and genome-wide association studies**. *Cancer biology & medicine* 2019, **16**(2):361–389.
18. Li J, Chang J, Zhu Y, Yang Y, Gong Y, Ke J, Lou J, Zhong R, Gong J, Xia X *et al*: **[Risk prediction of colorectal cancer with common genetic variants and conventional non-genetic factors in a Chinese Han population]**. *Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi* 2015, **36**(10):1053–1057.
19. Yan C, Zhu M, Ding Y, Yang M, Wang M, Li G, Ren C, Huang T, Yang W, He B *et al*: **Meta-analysis of genome-wide association studies and functional assays decipher susceptibility genes for gastric cancer in Chinese populations**. *Gut* 2020, **69**(4):641–651.
20. Tian J, Miao X, D. L: **Research progress of genetic risk prediction models for common malignant tumors in Chinese population**. *Biotechnology&Business* 2016, **6**(10):10–15.
21. Zhu M, Wang Y, Liu X, Wen X, Liang C, Tu J: **LncRNAs act as prognostic biomarkers in gastric cancer: A systematic review and meta-analysis**. *Frontiers in Laboratory Medicine* 2017:S2542364917300572.
22. Fattahi S, Kosari *cm* onfared M, Golpour M, Emami Z, Akhavan *km* iaki H: **LncRNAs as potential diagnostic and prognostic biomarkers in gastric cancer: A novel approach to personalized medicine**. *Journal of cellular physiology* 2019(168).
23. Esfandi F, Salehnezhad T, Taheri M, Afsharpad M, Hafez AA, Oskooei VK, Ghafouri-Fard S: **Expression assessment of a panel of long non-coding RNAs in gastric malignancy**. *Experimental and molecular pathology* 2020, **113**:104383.
24. Duan F, Song C, Shi J, Wang P, Ye H, Dai L, Zhang J, Wang K: **Identification and epidemiological evaluation of gastric cancer risk factors: based on a field synopsis and meta-analysis in Chinese population**. *Aging* 2021, **13**(17):21451–21469.
25. Peek RM, Jr., Blaser MJ: **Helicobacter pylori and gastrointestinal tract adenocarcinomas**. *Nature reviews Cancer* 2002, **2**(1):28–37.
26. Vaarhorst AA, Lu Y, Heijmans BT, Dolle ME, Bohringer S, Putter H, Imholz S, Merry AH, van Greevenbroek MM, Jukema JW *et al*: **Literature-based genetic risk scores for coronary heart disease: the Cardiovascular Registry Maastricht (CAREMA) prospective cohort study**. *Circulation Cardiovascular genetics*

2012, **5**(2):202–209.

27. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, Wang Q, Dennis J, Dunning AM, Shah M *et al*: **Prediction of breast cancer risk based on profiling with common genetic variants**. *Journal of the National Cancer Institute* 2015, **107**(5).
28. Song L, Liu A, Shi J, Molecular Genetics of Schizophrenia C: **SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics**. *Bioinformatics* 2019, **35**(20):4038–4044.
29. Lakeman IMM, Hilbers FS, Rodriguez-Girondo M, Lee A, Vreeswijk MPG, Hollestelle A, Seynaeve C, Meijers-Heijboer H, Oosterwijk JC, Hoogerbrugge N *et al*: **Addition of a 161-SNP polygenic risk score to family history-based risk prediction: impact on clinical management in non-BRCA1/2 breast cancer families**. *Journal of medical genetics* 2019, **56**(9):581–589.
30. Vachon CM, Scott CG, Tamimi RM, Thompson DJ, Fasching PA, Stone J, Southey MC, Winham S, Lindstrom S, Lilyquist J *et al*: **Joint association of mammographic density adjusted for age and body mass index and polygenic risk score with breast cancer risk**. *Breast Cancer Res* 2019, **21**(1):68.
31. D'Souza M, Schou M, Skals R, Weeke PE, Lee C, Smedegaard L, Madelaire C, Gerds TA, Poulsen HE, Hansen T *et al*: **Polygenic predisposition to breast cancer and the risk of coronary artery disease**. *International journal of cardiology* 2019, **291**:145–151.
32. McPherson R, Tybjaerg-Hansen A: **Genetics of Coronary Artery Disease**. *Circ Res* 2016, **118**(4):564–578.
33. Du Z, Hopp H, Ingles SA, Huff C, Sheng X, Weaver B, Stern M, Hoffmann TJ, John EM, Van Den Eeden SK *et al*: **A genome-wide association study of prostate cancer in Latinos**. *International journal of cancer Journal international du cancer* 2020, **146**(7):1819–1826.
34. Dong J, Buas MF, Gharahkhani P, Kendall BJ, Onstad L, Zhao S, Anderson LA, Wu AH, Ye W, Bird NC *et al*: **Determining Risk of Barrett's Esophagus and Esophageal Adenocarcinoma Based on Epidemiologic Factors and Genetic Variants**. *Gastroenterology* 2018, **154**(5):1273-1281 e1273.
35. Dai J, Lv J, Zhu M, Wang Y, Qin N, Ma H, He YQ, Zhang R, Tan W, Fan J *et al*: **Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations**. *The Lancet Respiratory medicine* 2019, **7**(10):881–891.

Tables

Table 1 Basic characteristic of individuals in case and control groups

Variables	Case (%)	Controls (%)	t/χ^2	*P*
Age (Mean ± SD)	57.64 ± 12.08	57.88 ± 11.50	0.355	0.552
Gender			0.000	1.000
Men	475 (71.97)	475 (71.97)		
Female	185 (28.03)	185 (28.03)		
Smoking status			12.80	<0.001
Non-smoker	398 (60.30)	460 (69.70)		
Yes	262 (39.70)	200 (30.30)		
Drinking status			9.25	0.002
Non-drinker	458 (69.39)	507(76.82)		
Drinker	202 (30.61)	153 (23.18)		
Family history			41.94	<0.001
No	587 (88.07)	646 (97.88)		
Yes	72 (11.93)	14 (2.12)		
Hp infection			7.74	0.005
No	89(37.08)	94(50.54)		
Yes	151(62.92)	92(49.46)		

SD: Standard deviation

Table 2
Association between the candidate lncRNA SNPs and risk of gastric cancer

SNP(rs#)	Per-allele		Heterozygous		Homozygous		Dominant model		Recessive mo
	OR(*95%CI*)	*P*	*OR*(*95%CI*)	P^a	*OR*(*95%CI*)	P^a	*OR*(*95%CI*)	P^a	*OR*(*95%CI*)
rs1859168	1.089(0.920,1.290)	0.321	0.389(0.275,0.496)	0.000	1.051(0.769,1.437)	0.755	0.649(0.492,0.857)	0.002	1.789(1.386,2
rs3815254	0.984(0.828,1.171)	0.860	1.012(0.778,1.316)	0.929	0.929(0.633,1.364)	0.706	0.993(0.774,1.275)	0.958	0.923(0.647,1
rs4784659	0.554(0.438,0.701)	0.000	0.420(0.313,0.565)	0.000	0.572(0.294,1.114)	0.100	0.438(0.331,0.579)	0.000	0.710(0.367,1
rs579501	0.714(0.557,0.917)	0.008	0.729(0.542,0.981)	0.037	0.517(0.224,1.191)	0.121	0.705(0.530,0.939)	0.017	0.555(0.242,1
rs77628730	1.261(1.046,1.521)	0.015	1.206(0.936,1.554)	0.147	1.807(1.085,3.011)	0.023	1.273(0.997,1.624)	0.053	1.656(1.008,2
rs6989575	1.030(0.870,1.219)	0.733	1.200(0.902,1.595)	0.211	1.004(0.703,1.433)	0.984	1.141(0.871,1.496)	0.338	0.892(0.658,1
rs7816475	1.191(0.960,1.478)	0.112	1.435(1.097,1.878)	0.008	0.868(0.451,1.672)	0.673	1.358(1.049,1.757)	0.020	0.776(0.405,1
rs6470502	0.505(0.406,0.628)	0.000	0.329(0.244,0.445)	0.000	0.629(0.387,1.023)	0.062	0.382(0.292,0.501)	0.000	0.840(0.521,1
rs1518338	1.084(0.890,1.320)	0.421	1.355(1.051,1.747)	0.019	0.635(0.347,1.163)	0.141	1.251(0.979,1.598)	0.074	0.561(0.309,1
rs2867837	0.948(0.766,1.173)	0.625	0.697(0.524,0.927)	0.013	1.582(0.955,2.622)	0.157	0.827(0.637,1.073)	0.153	1.742(1.057,2
rs12494960	2.616(2.122,3.226)	0.000	2.566(1.967,3.347)	0.000	7.672(3.790,15.530)	0.000	2.897(2.241,3.744)	0.000	5.392(2.681,1
rs74798803	0.966(0.795,1.174)	0.728	0.992(0.772,1.274)	0.947	0.844(0.463,1.538)	0.580	0.976(0.765,1.245)	0.845	0.847(0.469,1
rs7818137	1.198(1.012,1.417)	0.036	1.581(1.169,2.139)	0.003	1.432(0.991,2.069)	0.056	1.539(1.152,2.056)	0.004	1.036(0.768,1
rs550894	1.129(0.934,1.364)	0.209	1.242(0.964,1.601)	0.093	1.274(0.764,2.124)	0.353	1.264(0.977,1.591)	0.077	0.865(0.526,1
rs3825071	1.475(1.161,1.873)	0.001	1.687(1.278,2.227)	0.000	1.136(0.405,3.191)	0.808	1.654(1.260,2.171)	0.000	0.986(0.352,2
rs580933	0.980(0.807,1.191)	0.843	1.160(0.897,1.500)	1.160	0.808(0.486,1.343)	0.411	1.099(0.860,1.405)	0.450	0.762(0.463,1
rs7943779	1.537(1.194,1.978)	0.001	1.615(1.215,2.147)	0.001	2.078(0.484,8.918)	0.325	1.627(1.228,21.56)	0.001	1.840(0.429,7
rs911157	1.741(1.192,2.542)	0.004	1.651(1.099,2.480)	0.016	1.869(0.157,22.253)	0.621	1.656(1.107,2.477)	0.014	1.771(0.148,2
rs16981280	0.756(0.636,0.899)	0.002	0.677(0.519,0.884)	0.002	0.539(0.364,0.798)	0.002	0.646(0.500,0.833)	0.000	0.677(0.473,0
rs2273534	0.919(0.777,1.087)	0.324	1.073(0.794,1.449)	0.646	0.902(0.631,1.291)	0.574	1.019(0.765,1.358)	0.896	0.859(0.643,1
rs957313	1.040(0.790,1.370)	0.779	1.110(0.810,1.520)	0.518	1.203(0.392,3.687)	0.747	1.115(0.820,1.516)	0.487	1.180(0.385,3

P^a value of logistic regression analysis adjusted for age, gender, smoking, drinking and family history of tumors in first-degree relatives

Table 3
Association between the Common SNPs and risk of gastric cancer

SNP(rs#)	Per-allele		Heterozygous		Homozygous		Dominant model		Recessive
	OR(*95%CI*)	*P*	*OR*(*95%CI*)	P^a	*OR*(*95%CI*)	P^a	*OR*(*95%CI*)	P^a	*OR*(*95%C*
rs861539	0.729(0.524,1.013)	0.059	0.760(0.532,1.086)	0.132	0.343(0.035,3.399)	0.360	0.702(0.493,1.00)	0.050	0.368(0.0
rs2294008	1.228(1.038,1.452)	0.017	1.117(0.883,1.412)	0.356	1.560(1.046,2.326)	0.029	1..252(1.002,1.564)	0.048	1.444(0.9
rs731236	0.927(0.658,1.305)	0.663	0.925(0.640,1.337)	0.677	0.536(0.048,5.952)	0.611	0.900(0.624,1.296)	0.571	0.242(0.0
rs25487	1.168(0.980,1.392)	0.082	1.294(1.023,1.363)	0.032	1.033(0.658,1.623)	0.888	1.241(0.992,1.551)	0.059	0.915(0.5
rs751402	1.173(0.988,1.379)	0.053	1.259(0.995,1.593)	0.055	1.256(0.859,1.835)	0.240	1.270(1.014,1.590)	0.037	1.091(0.7
rs1801133	1.233(1.057,1.440)	0.003	1.262(0.984,1.619)	0.066	1.474(1.070,2.032)	0.018	1.310(1.039,1.653)	0.023	1.256(0.9
rs1799782	1.154(0.980,1.358)	0.086	1.254(0.994,1.582)	0.056	1.711(1.124,2.606)	0.012	1.348(1.078,1.685)	0.009	1.632(1.0
rs763780	1.356(1.070,1.720)	0.012	1.432(1.090,1.881)	0.010	1.449(0.529,3.965)	0.471	1.386(1.061,1.811)	0.017	1.289(0.4
rs8193036	1.175(0.999,1.382)	0.052	1.233(0.975,1.560)	0.081	1.307(0.896,1.908)	0.165	1.267(1.013,1.585)	0.038	1.076(0.7
rs4072037	0.692(0.553,0.866)	0.001	0.635(0.487,0.827)	0.001	0.778(0.338,1.788)	0.554	0.666(0.515,0.860)	0.002	0.852(0.3
rs2274223	1.209(1.014,1.442)	0.035	1.083(0.856,1.370)	0.506	1.690(1.055,2.707)	0.029	1.163(0.929,1.455)	0.188	1.625(1.0
rs2275913	1.199(1.027,1.399)	0.022	1.037(0.806,1.335)	0.776	1.349(0.987,1.844)	0.060	1.143(0.905,1.444)	0.263	1.273(0.9
rs1799929	0.863(0.552,1.349)	0.518	0.866(0.544,1.378)	0.543	–		0.807(0.506,1.289)	0.370	0.753(0.5
rs20417	1.530(1.101,2.125)	0.011	1.557(1.093,2.219)	0.014	1.372(0.115,16.312)	0.802	1.533(1.078,2.180)	0.017	1.810(0.1
rs1800896	1.023(0.779,1.356)	0.846	1.054(0.776,1.431)	0.738	0.615(0.145,2.617)	0.510	0.925(0.682,1.255)	0.617	0.480(0.1
rs13361707	1.193(1.024,1.391)	0.024	0.750(0.575,0.976)	0.032	0.674(0.488,0.933)	0.017	0.755(0.588,0.969)	0.028	0.827(0.6
rs3773651	0.715(0.547,0.936)	0.014	0.703(0.516,0.956)	0.025	0.815(0.279,2.378)	0.708	0.654(0.483,0.884)	0.006	0.661(0.2
rs1799930	1.511(1.270,1.798)	<0.001	1.572(1.241,1.991)	<0.001	1.649(1.060,2.563)	0.026	1.663(1.329,2.082)	<0.001	1.632(1.0
GSTM1	–	–	–	–	1.362(1.089,1.703)	0.007	–	–	–
GSTT1	–	–	–	–	1.012(0.808,1.267)	0.918	–	–	–

P^a value of logistic regression analysis adjusted for age, gender, smoking, drinking and family history of tumors in first-degree relatives

Table 4
Comparison of NRI and IDI between different risk prediction models

Model comparison	NRI	*Z*	*P*	IDI	*Z*	*P*
wGRS vs PRS	0.108	4.122	<0.001	0.046	187.569	<0.001
wGRS vs wGRS+lncRNAs	0.174	6.351	<0.001	0.087	285.556	<0.001
wGRS vs PRS+lncRNAs	0.036	1.833	0.067	0.023	135.237	<0.001
wGRS+lncRNAs vs PRS+lncRNAs	0.138	5.084	<0.001	0.064	208.313	<0.001

IDI: Integrated Discrimination Improvement; NRI: Net Reclassification Improvement.

Table 5
Comparison of AUC, AIC and BIC of different risk prediction models

Model and variables	AUC(95%CI)	AIC	BIC
wGRS	0.634(0.605,0.664)	1878.857	1889.227
wGRS+lncRNAs	0.707(0.679,0.734)	1836.195	1846.566
PRS	0.679(0.650,0.708)	1814.786	1824.157
PRS+lncRNAs	0.730(0.703,0.757)	1745.085	1755.456
PRS+*Hp* infection	0.724(0.659,0.789)	315.668	322.654
PRS+Drinking	0.730(0.678,0.783)	640.100	648.371
PRS+Smoking	0.657(0.607,0.706)	462.319	470.263
PRS+Smoking+Drinking	0.719(0.656,0.783)	332.783	339.802
PRS+Smoking+Drinking+*Hp* infection	0.774(0.767,0.871)	118.284	123.371
PRS+lncRNAs+Smoking+Drinking+*Hp* infection	0.779(0.682,0.875)	117.228	122.314
AUC: Area Under Curve; AIC: Akaike Information Criterion;BIC, Bayesian Information Criterion.			

Figures

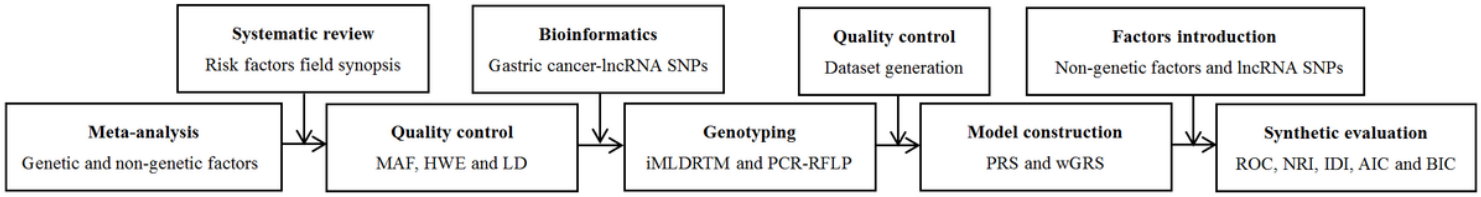

Figure 1

Flowchart for the study design. lncRNAs, long non-coding RNAs; MAF, Minor allele frequency; HWE, Hardy-Weinberg Equilibrium; LD, Linkage disequilibrium; iMLDR, Improved Multiplex Ligation Detection Reaction; PCR-RFLP, Polymerase chain reaction restriction fragment length polymorphism; PRS, Polygenic risk scores; wGRS, weighted genetic risk scores; ROC, Receiver operating characteristic; NRI, Net reclassification improvement; AIC, Akaike information criterion; BIC, Bayesian information criterion.

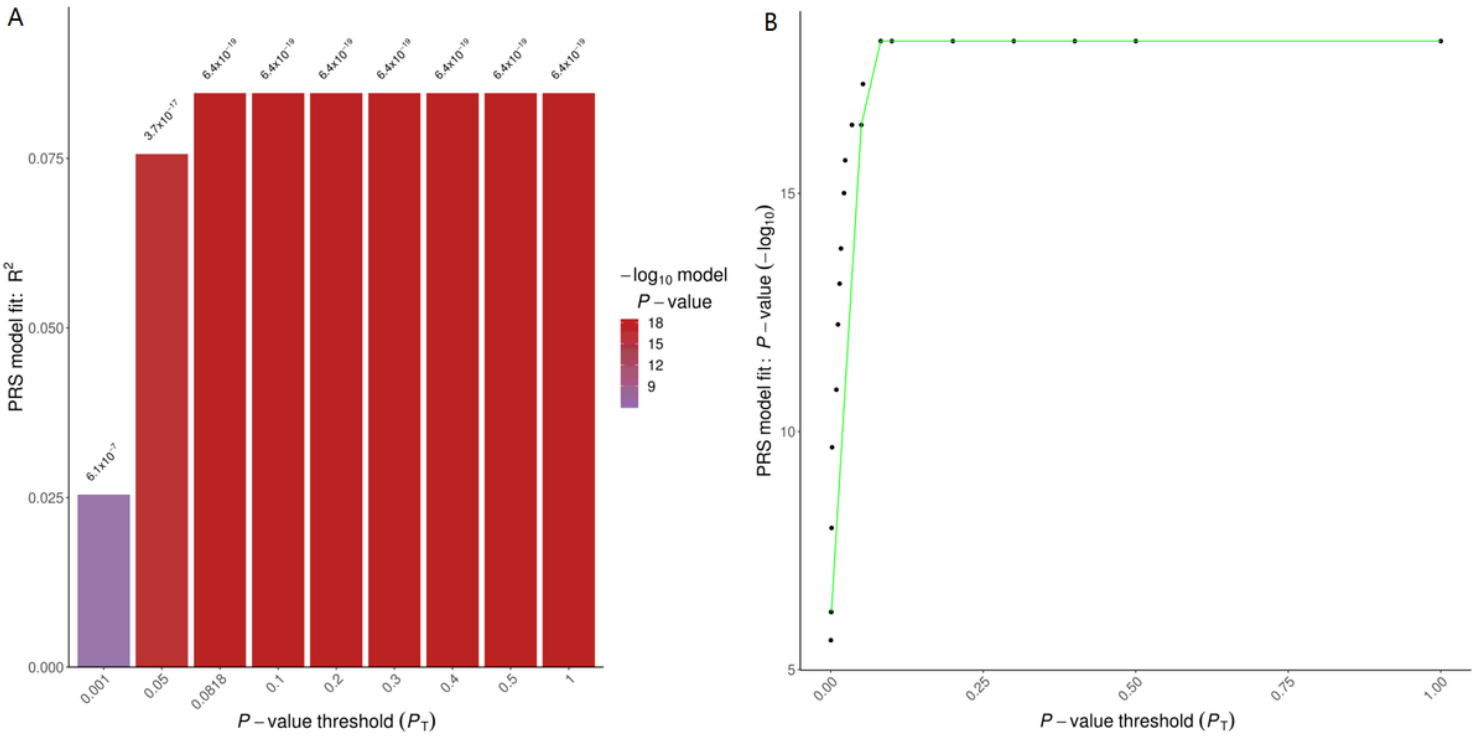

Figure 2

Predictive P-value threshold (PT) and phenotypic variation interpretation bar chart of gastric cancer (A) and high resolution plot of PRS P value threshold (PT) and model goodness of fit.

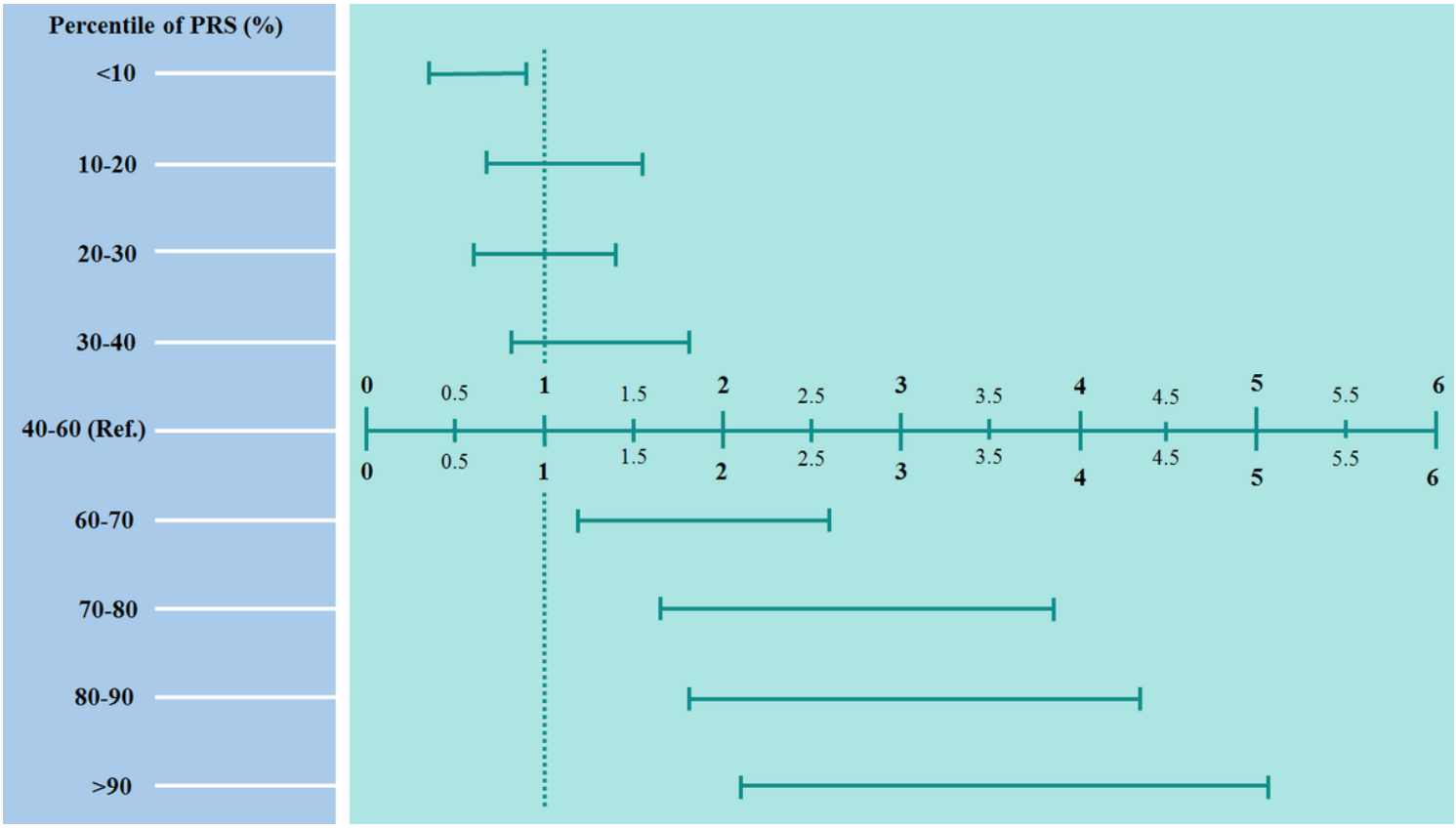

Figure 3

Regression analysis predicts the risk quantile plot of the PRS phenotype.

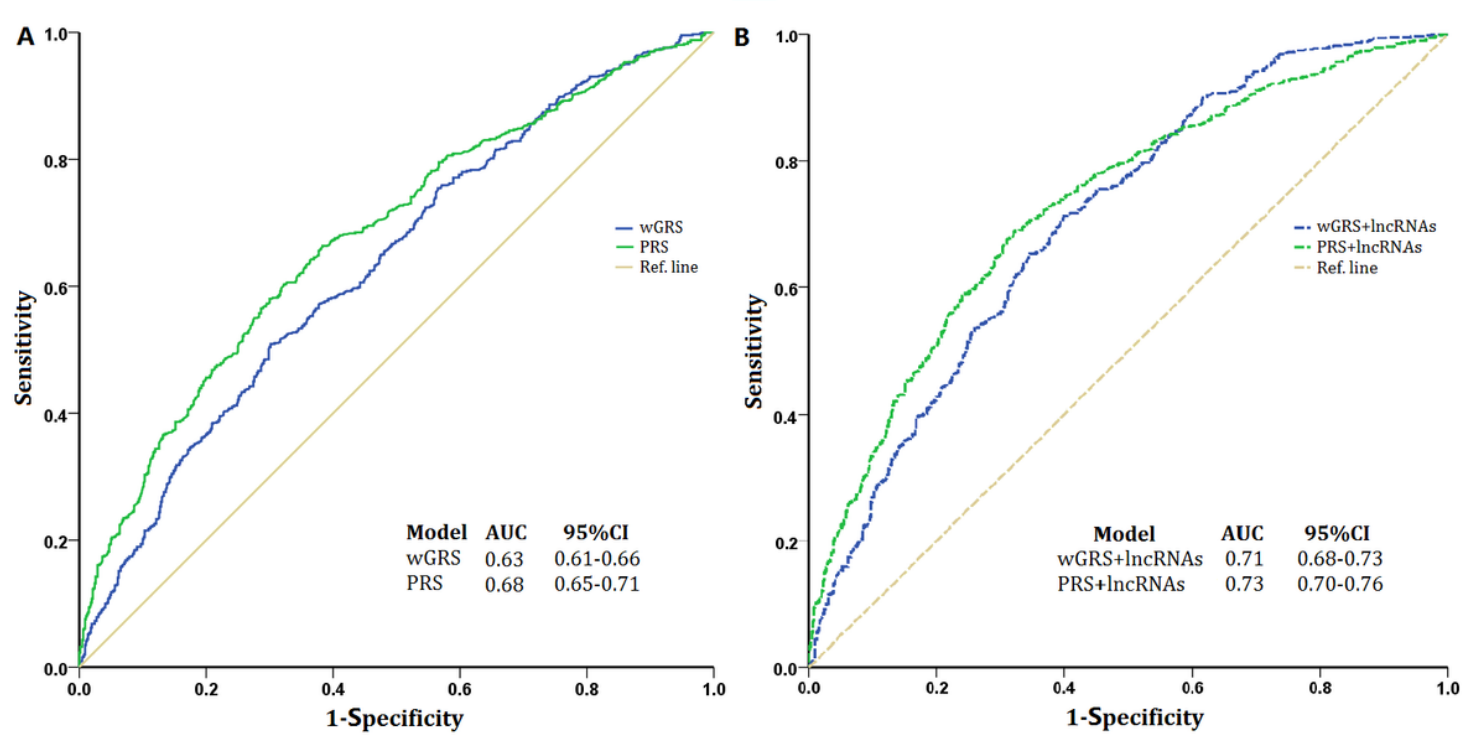

Figure 4

ROC curve of different genetic risk factors prediction models for gastric cancer.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx