The Construction of Polygenic Risk Scores for Breast Cancer Based on LightGBM and Multiple Omics Data

Baoshan Ma (mabaoshan@dlmu.edu.cn)
Dalian Maritime University https://orcid.org/0000-0003-1629-1968
Jianqiao Pan
Dalian Maritime University
Xiaoyu Hou
Dalian Maritime University
Chongyang Li
Dalian Maritime University
Tong Xiong
Dalian Maritime University
Yi Gong
Dalian Maritime University
Fengju Song
Tianjin Medical University Cancer Institute and Hospital: Tianjin Tumor Hospital

Research Article

Keywords: Breast cancer, Polygenic risk scores, Multiple omics data, LightGBM, Diagnosis, Prognosis

Posted Date: April 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-438740/v1

The Construction of Polygenic Risk Scores for Breast Cancer Based on LightGBM and Multiple Omics Data

Baoshan Ma[1,*], Jianqiao Pan[1], Xiaoyu Hou[1], Chongyang Li[1], Tong Xiong[1], Yi Gong[1], Fengju Song[2,*]

[1] School of Information Science and Technology, Dalian Maritime University, Dalian, 116026, China

[2] Department of Epidemiology and Biostatistics, Key Laboratory of Molecular Cancer Epidemiology, Tianjin, National Clinical Research Center of Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin, 300060, China

***Corresponding author:**

E-mail addresses: mabaoshan@dlmu.edu.cn (Baoshan Ma), songfengju@163.com (Fengju Song)

Tel: +86 18624392829 (Baoshan Ma), +86 13164066716 (Fengju Song)

Abstract

Background: Breast cancer accounts for a large proportion of cancer-related deaths in women. Polygenic risk score (PRS) derived from single nucleotide polymorphisms (SNP) data can evaluate the individual-level genetic risk of breast cancer and has been widely applied for risk stratification. However, standalone SNP data used for PRS may not provide satisfactory prediction accuracy. Additionally, current PRS models based on linear regression have insufficient power to leverage non-linear effects from thousands of associated SNPs.

Methods: In this study, the multiple omics data (DNA methylation data, miRNA data, mRNA data and lncRNA data) and clinical data of breast invasive carcinoma (BRCA) were collected from The Cancer Genome Atlas (TCGA). First, we developed a novel PRS model utilizing single omic data and a machine learning algorithm (LightGBM). Subsequently, we built a combination model of PRS derived from each omic data to explore whether multiple omics data can further improve the prediction accuracy of PRS. Finally, we performed association analysis and prognosis prediction of breast cancer to evaluate the utility of the PRS generated by our method.

Results: Our PRS model based on single omic data and LightGBM algorithm achieved better predictive performance than the linear models and other machine learning models. Moreover, the combination of the PRS derived from each omic data can efficiently strengthen prediction accuracy. The analysis of prevalence and the associations of the PRS with phenotypes including case-control and cancer stage status indicated that the risk of breast cancer increases

with the increases of PRS. The survival analysis also suggested that PRS for the cancer stage is an effective prognostic metric of breast cancer patients.

Conclusion: Our proposed model expanded the current definition of PRS from standalone SNP data to multiple omics data and outperformed the state-of-the-art PRS models, which may provide a powerful tool for diagnostic and prognostic prediction of breast cancer.

Keywords: Breast cancer, Polygenic risk scores, Multiple omics data, LightGBM, Diagnosis, Prognosis

1. Introduction

Breast cancer is the most frequently diagnosed cancer in women worldwide[1]. In 2020, there were over 2 million new cases reported[2]. The establishment of effective prevention and treatment measures is essential to prevent breast cancer occurrence and reduce breast cancer mortality. Although carriers of BRCA1 and BRCA2 gene mutations confer a high risk of breast cancer, these gene mutations can be found in only a small part of breast cancer patients[3]. In recent years, genome-wide association study (GWAS) identified multiple high frequency and low penetrance susceptibility variants of breast cancer[4]. The accumulation effects of these susceptibility variants can be summarized as a polygenic risk score (PRS). In recent years, researchers have developed several PRS models for breast cancer by using a large amount of single nucleotide polymorphisms (SNPs) data. These studies maintained the PRS to be an effective and reliable predictor of breast cancer risk that may provide screening and prevention strategies[5-9].

However, the PRS calculated using SNPs data can only assess the genetic risk of an individual, while ignoring the influence of the external environmental exposure on gene expression. With the development of high-throughput omics technology, a large number of related studies based on genomics and transcriptomics emerged[10, 11]. These high-throughput molecular markers can dynamically reflect the comprehensive effects of genetic background, environmental exposure and lifestyle habits individually[12-14]. The analyses of multiple omics data may lead to new insights into diagnosis and prognosis of

81 breast cancer[15]. In addition, in the standard approach of PRS, the effect sizes of the genetic
82 variants are usually estimated in linear statistical models[16-19]. However, linear statistical
83 model has some limitations and only be applied when specific requirements are satisfied[20].
84 Advanced machine learning (ML) models[21, 22] such as LightGBM can account for
85 non-linear relationships among large-scale variables and have an increasing trend on the
86 applications for breast cancer research. Using these ML models may further improve the
87 prediction accuracy (PA) of PRS.

88 Here, we used multiple omics data and LightGBM model to construct a novel PRS for breast
89 cancer. The results illustrated that our proposed method outperforms traditional linear models
90 and other state-of-the-art ML models and can effectively predict individual risk of breast
91 cancer.

92 # 2. Methods

93 ## 2.1 Data collection

94 The datasets in this study were downloaded from The Cancer Genome Atlas (TCGA) project.
95 Now, all TCGA data are accessible without limitations in publications or presentations
96 according to the posted announcement from the TCGA website[23]. We collected four kinds of
97 omics datasets on breast invasive carcinoma (BRCA), including DNA methylation data
98 (Illumina Infinium Human DNA Methylation 450K; Level 3) measured from 782 tumor tissues
99 and 96 normal tissues (Paracancerous tissue), miRNA-seq data (IlluminaHiSeq_miRNASeq;
100 Level 3) measured from 1078 tumor tissues and 104 normal tissues, mRNA expression

(Illumina mRNA-seq; Level 3) measured from 1102 tumor tissues and 113 normal tissues, lncRNA expression (Illumina lncRNA-seq; Level 3) measured from 1102 tumor tissues and 113 normal tissues. We also collected the stage of tumor for the BRCA patients, including stage I, stage II, stage III, and stage IV. According to the literature[24], the annotation of stage I and II were labelled as early-stage, stage III and IV as late-stage. For BRCA patients, most of the individuals are white, and a small number of individuals are black or African American and Asian. Although the female patients of breast cancer account for the majority, there are also 5 male patients in the BRCA datasets. The ages of volunteers used in our study range from 26 to 90. Table 1 and Supplementary Figure 1 shows the number of sampels in BRCA datasets and the detail of basic demographic characteristics, respectively.

Table 1. The description of datasets used in this study.

Omic type	Total of early-stage and late-stage tumor samples		Total of tumor samples	Total of normal samples	Total of biological variables
DNA methylation	Early-stage	562	782	96	14797
	Late-stage	209			
miRNA	Early-stage	790	1078	104	360
	Late-stage	264			
mRNA	Early-stage	800	1102	113	16499
	Late-stage	267			
lncRNA	Early-stage	800	1102	113	5382
	Late-stage	267			

The total of tumor samples is not equal to the sum of the early-stage and the late-stage samples, because some tumor samples have unknown breast cancer stage.

2.2 Data pre-processing

For DNA methylation, we retained the CpG sites that most negatively correlated with gene expression according to Firehose[25] and removed CpG sites with missing value to ensure the quality of the datasets[26]. For miRNA, mRNA, and lncRNA, two steps were performed to deal with the missing values in the datasets[27]. First, the probes were excluded if there is missing value in more than 20% of samples. Second, all data were normalized by Min-Max scaling to map the range from 0 to 1. For convenience, CpG sites of DNA methylation and probes of miRNA, mRNA and lncRNA are collectively referred to as biological variables. Table 1 also shows the the summary of biological variables.

2.3 Construction of polygenic risk scores

2.3.1 Overview of PRS model

According to the different phenotypes, we proposed to utilize multiple omics data and breast cancer status to construct two kinds of PRS models. The first phenotype only contains the normal samples (control) and tumor samples (case), which were labelled 0 and 1, respectively. The second phenotype contains the normal samples, early-stage and late-stage tumor samples, which were labelled 0, 1 and 2, respectively. We defined the above-mentioned two PRS models as PRS for case-control status and PRS for cancer stage status. The PRS can evaluate the individual risk of breast cancer and may improve the diagnosis of breast cancer. Moreover, since recent studies found the stage of cancer is highly associated with the prognosis[28], accurate construction of PRS for cancer stage status may facilitate the prediction of breast cancer prognosis. The framework of this study is shown in Figure 1.

2.3.2 PRS based on LightGBM

LightGBM is an ensemble model of classification and regression trees (CART)[29], in which each step generates a basic CART model and adds it to the overall model. The PRS models based on LightGBM were built using a training dataset to predict PRS in a testing dataset. We defined each omic dataset $D_1 = \{(\mathbf{X}_i, y_i)\}(|\ D_1\ | = n_1, \mathbf{X}_i \in \mathbf{R}^m, y_i \in \mathbf{R})$ as training dataset, where $\mathbf{X}_i$ represents a matrix containing n_1 samples and m biological variables, y_i is the corresponding outcome (phenotype). Let $\hat{y}_i$ be the prediction of y_i. $D_2 = \{(\mathbf{X}_i^*, y_i^*)\}(|\ D_2\ | = n_2, \mathbf{X}_i^* \in \mathbf{R}^m, y_i^* \in \mathbf{R})$ was the testing dataset, where $\mathbf{X}_i^*$ represents a matrix containing n_2 samples and m biological variables, y_i^* denotes the PRS. We used T additive CART models to predict the PRS in the training dataset.

$$\hat{y}_i = \sum_{t=1}^{T} f_t(\mathbf{X}_i), f_t \in F \quad (1)$$

where $f_t(\mathbf{X}_i)$ corresponds to an independent CART model and F is the space of CART models. To learn the set of CART used in the PRS model, we minimize the following objective function.

$$L(f_t) = \sum_{i=1}^{n_1} l(y_i, \hat{y}_i) + \gamma K + \frac{1}{2}\lambda \sum_{k=1}^{K} w_k^2 \quad (2)$$

Here $l(y_i, \hat{y}_i)$ is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and true phenotype y_i. The K and w_k respectively represent the number and value of leaf nodes in each CART model, γ and λ are constant coefficients.

In general setting, the second-order approximation can be utilized to quickly optimize the objective function.

$$L(f_t) = \sum_{k=1}^{K}[(\sum_{i \in I_k} g_i)w_k + \frac{1}{2}(\sum_{i \in I_k} h_i + \lambda)w_k^2] + \gamma K \quad (3)$$

where g_i and h_i are the first and second order gradient statistics of the loss function. $I_k = \{i \mid q(\mathbf{X}_i) = k\}$ was defined as the instance set of leaf node. LightGBM used two techniques including gradient-based one-side sampling and exclusive feature bundling to estimate the information gain in a high speed[21]. The structure and value of each CART model can be determined by the information gain. Thus, we generated the PRS model consisting of T additive CART models. For the samples in a testing dataset, PRS y_i^* can be calculated by applying $\mathbf{X}_i^*$ to the PRS model. We also provided an automatic python program based on our proposed models to obtain PRS, which is available for downloading from GitHub website[30].

2.3.3 PRS based on linear model and other ML model

To evaluate the predictive performance of LightGBM objectively, we applied the traditional linear model and other state-of-the-art ML models to construct PRS. The traditional linear model contains minimax concave penalty (MCP)[17, 31], least absolute shrinkage and selection operator (LASSO)[18, 32] and elastic net[33]. The ML model contains support vector regression (SVR)[34]. Here, we compared the PRS methods that only utilize omics data, without considering the methods that use GWAS summary statistics, such as

LDpred[35], Lassosum[36] and so on. Similar to the PRS method based on LightGBM model, we used each omic dataset as the input of these models, and the corresponding phenotypes as the output.

2.3.4 Model training and evaluation

To ensure the robustness and stability of the model, we trained and evaluated the proposed PRS model by 5-fold cross validation. This procedure divided each omic dataset into five subsets. In each fold, one of the five subsets was used as the testing dataset and the other four subsets were put together to form a training dataset. We applied bayesian optimization[37] and 3-fold inner cross validation to optimize the hyper-parameters of the PRS model in each training dataset. Specifically, for LASSO, we optimized the parameter "alpha". For MCP, we adjusted regularization parameter "labmda". For elastic net, the parameter "alpha" and "l1_ratio" were optimized. For SVR, we choose "rbf kernel" and optimized the regularization parameter "C". For LightGBM, the optimized parameters were "num_leaves", "n_estimators", "learning_rate", "max_depth", "max_bin", "min_split_gain", "subsample", "subsample_freq", "colsample_bytree", "min_child_sample", "min_child_weight", "reg_alpha", "reg_lambda". Finally, we obtained the PRS of each testing dataset which was predicted by the model with the optimized parameters. Each PRS was standardized based on its mean and standard deviation. The predictive performance of PRS model was evaluated by square of the Pearson correlation coefficient (R^2).

$$R^2 = \left(\frac{Cov(Y,\hat{Y})}{\sqrt{Var(Y)Var(\hat{Y})}}\right)^2 \quad (4)$$

where $Cov(Y,\hat{Y})$ represents the covariation of true phenotype and predicted PRS, $Var(Y)$ is the variance of true phenotype, and $Var(\hat{Y})$ is the variance of predicted PRS. In addition, for case-control status, we can also evaluate the predictive performance by the area under the receiver operating characteristic curve (AUC).

2.4 Combination model of PRS

To further improve the predictive performance of PRS, we utilized the PRS based on each omic dataset to construct a new combination model[38, 39]. We first matched a common dataset from the four kinds of omics datasets for BRCA. In the common dataset of case-control status, there are 786 tumor samples and 75 normal samples. In the common dataset of cancer stage status, there are 553 early-stage samples, 205 late-stage samples and 75 normal samples. Next, we used the PRS based on four kinds of omics datasets as new biological variables for the combination model. Then, we build the combination model using the LightGBM model. Bayesian optimization was applied to adjust hyper-parameters and 5-fold cv was used to evaluate the overall predictive performance. The framework of the combination model is shown in Supplementary Figure 2.

3. Results

3.1 Predictive performance of PRS based on multiple omics data

We first compared our prediction model to existing PRS methods and other ML methods for case-control status. Figure 2a shows the results of these PRS methods on four kinds of omics datasets. We observed that elastic net achieves the best performance in traditional linear models. The R^2 of SVR is 3.3%, 7.7% and 0.5% higher than elastic net on DNA methylation, miRNA and lncRNA datasets and 3.1% lower than elastic net on mRNA dataset. The R^2 of our proposed model improved by 8.3%, 14.8%, 5.1% and 7.2% than elastic net on four kinds of omics datasets. Overall, our model outperformed other models and mRNA data exhibited better performance than other omics data.

Next, we applied our proposed model and other PRS methods for cancer stage status. Compared with the case-control status, this phenotype contains normal and two stage statuses of breast cancer. Thus, the predictive performance of PRS for cancer stage status is not as good. Nevertheless, the present results are consistent with the PRS for case-control status. According to the comparison results of our proposed model with other PRS methods (Figure 2b), the LightGBM model performs the best predictive performance, outscoring other PRS methods on four kinds of omics datasets. The PRS based on LightGBM obtains the R^2 of 0.405, 0.371, 0.437 and 0.407, respectively. Compared with the elastic net with the highest PA in the linear models, the R^2 of LightGBM improved by 12.8%, 20.9%, 9.8% and 12.4%, respectively. Compared with SVR, the R^2 of our proposed model improved by 10.4%, 14.4%, 10.6% and 3.3%, respectively. Moreover, the results showed mRNA data obtained better results than other omics data.

3.2 Predictive performance of PRS based on combination model

In this part, we evaluated the performance of combination model of PRS. Figure 2c shows the results of PRS models based on four types of omics datasets and combination model in the common samples. For four kinds of omics datasets, although the PA in the common samples has decreased, we found that PRS based on mRNA still obtained the best PA. For case-control and cancer stage status, the R^2 of combination model were 0.932 and 0.397, respectively. Compared with the PRS model based on mRNA dataset, the R^2 of combination model improved by 5.1% and 2.8%, respectively. Thus, the combination of four types of molecular data can achieve better results of PRS for case-control and cancer stage status.

3.3 Prevalence of breast cancer

Exploring the prevalence of different PRS strata has a positive impact on the prevention and treatment of breast cancer[40]. The main goal of this part is to analyze the risk stratification of case-control status. Thus, we divided the common samples into 10 strata of increasing PRS from the combination model and calculated the prevalence of each stratum (Figure 3).

Across the common samples, we observed that the prevalence is about 10% in the first stratum then upgrades to 100% in the second stratum and remains steady afterwords. The prevalence changes significantly at one stratum because our proposed method achieved relatively accurate prediction of breast cancer risk for case-control status. The trend plot of the prevalence also indicated that individuals with high-PRS strata have greater breast cancer risk than the individuals with lower-PRS strata.

3.4 Associations between PRS and breast cancer risk

We investigated the relationship of PRS with different phenotypes of breast cancer in this section (Table 2). For case-control status, the association of PRS was evaluated in predicted results from the combination model by Logistic regression. We observed that PRS was associated with occurrence risk of breast cancer (odds ratio (OR) = 18.48; 95% confidence interval (CI): 9.60-35.55; $P = 2.46 \times 10^{-18}$), suggesting that per one standard deviation increase in PRS is associated with 170% risk increase of breast cancer. For cancer stage status, we performed a multinomial Logistic regression model to evaluate the association of PRS and set the normal sample as the reference group. The PRS was associated with early-stage breast cancer risk (OR = 21.05; 95%CI: 10.26-43.19; $P = 9.63 \times 10^{-17}$) and late-stage breast cancer risk (OR = 46.62; 95%CI: 19.72-110.25; $P=2.14 \times 10^{-18}$). The results indicated higher PRS is associated with a significantly increased risk for early-stage and late-stage breast cancer.

Table 2. Associations between PRS and breast cancer risk.

Phenotype		OR	95%CI	P-value
The PRS for case-control status		18.48	9.60 – 35.55	2.46×10^{-18}
The PRS for cancer stage status	Early-stage	21.05	10.26 – 43.19	9.63×10^{-17}
	Late-stage	46.62	19.72 – 110.25	2.14×10^{-18}

PRS: Polygenic risk score; OR: Odds ratio; CI: Confidence intervals.

3.5 Prognosis prediction of breast cancer

We explored whether the PRS for cancer stage status can effectively assess the prognosis of patients. According to the predicted results of tumor samples using combination model, we firstly divided 758 patients of breast cancer into high-risk and low-risk groups based on the

50th percentile of PRS. Next, we utilized the survival time and the status at the end of their survival time for each patient to generate Kaplan-Meier curves (KM curve)[41]. We observed that high-risk patients had statistically significantly worse prognosis (Figure 4). The results showed the PRS for cancer stage may provide an effective prognostic tool of breast cancer patients.

4. Discussion

In our study, first of all, we have developed a novel PRS method for breast cancer using multiple omics data and LightGBM model. For case-control and cancer stage status, we showed that the proposed method had better prediction performance than existing traditional linear models and state-of-art ML models using multiple omics data. Meanwhile, the prediction results of 5-fold-cv demonstrated the robustness and reliability of our proposed method. Second, the combination of PRS further improved the predictive performance for breast cancer. Finally, by analyzing the trend of prevalence and associations between PRS and breast cancer risk, the results bolstered the clinical understanding and application for breast cancer PRS. In addition, we also found that our PRS models for cancer stage status can improve the prognosis prediction of breast cancer patients.

Most of the previous PRS studies focused on the analysis of individual-level genotype data (SNPs) using linear models. For example, Mavaddat *et al.* utilized PRS derived from 313 SNPs in 69 studies of the Breast Cancer Association Consortium (BCAC) to predict the breast cancer risk and the AUC was 0.63[9]. Khera *et al.* derived a PRS based on 5218 SNPs in the

UK Biobank and the AUC was 0.68[8]. Although these studies obtain individual-level genetic risk of breast cancer, the current PA still maintains at low level. In case-control status of our study, we obtained the AUC of 0.99 from the combination model. Thus, the PRS based on multiple omics data and LightGBM not only improved the risk of predicting breast cancer, but also expanded the current definition of PRS from SNP data to genomics and transcriptomics data.

Our proposed method outperforms the other breast cancer PRS for two main reasons. First,the LightGBM model fits all biological variables simultaneously using gradient boosting tree, especially high-dimensional data such as multiple omics data[21], while linear model such as MCP, LASSO only utilizes marginal variables to construct PRS. In addition, the LightGBM model takes advantage of ensemble learning, which helps to minimize the main causes of error in ML model such as noise, bias and variance than a single model[42]. Second, as representative of genomics data, DNA methylation can be modulated by physiological and environmental exposures and provided biomarkers for diagnosis and prognosis for cancer[43, 44]. Transcriptomics data including miRNA, mRNA, and lncRNA reveals the transcription and regulation mechanism of large-scale genes, which play an important role in determining the mechanism and treatment of cancer[45, 46]. Compared to the individual-level genotype data, using multiple omics data to construct breast cancer PRS considered the interaction of genetic and environmental factors, and thus can provide higher PA.

Although our PRS methods provide powerful predictive performance, they have some limitations. First, the LightGBM model has more hyper-parameters than traditional linear models such as MCP, LASSO and elastic net. Thus, we need more time to train the proposed model. We applied multithreading technology to effectively utilize computing resources and correspondingly reduced some running time. Second, the sample size of breast cancer from TCGA is relatively small compared to large-scale Genome-wide association studies data. In addition, there are significantly more tumor samples than normal samples in our study. Imbalanced datasets significantly compromise the performance of most standard learning algorithms, because these models assume the balanced class distributions. Third, this study lacks independent validation datasets, because it is very difficult to collect multiple omics data including DNA methylation, miRNA, mRNA and lmcRNA of case-control and cancer stage status. Thus, we employed 5-fold cross validation to strengthen the robustness and stability of our proposed models. In the future, we will consider applying our PRS model to analyze breast cancer with other phenotypes by using larger and balanced multiple omics datasets.

5. Conclusions

In conclusion, we proposed a novel PRS model in two kinds of breast cancer phenotypes by using multiple omics data and LightGBM. The results demonstrated our model improved the PA of current PRS methods indeed and may provide an effective diagnosis and prognosis tool for breast cancer.

Abbreviations

GWAS: Genome-wide association study; PRS: Polygenic risk score; SNPs: Single nucleotide polymorphisms; ML: Machine learning; PA: Prediction accuracy; TCGA: The Cancer Genome Atlas; BRCA: Breast invasive carcinoma; CART: Classification and regression trees; MCP: Minimax concave penalty; LASSO: Least absolute shrinkage and selection operator; SVR: Support vector regression; AUC: Area under the receiver operating characteristic curve; OR: Odds ratio; CI: Confidence interval; KM: Kaplan-Meier; BCAC: Breast Cancer Association Consortium

Declarations

Ethics approval and consent to participate

All TCGA data are accessible without limitations in publications or presentations according to the posted announcement from the TCGA website.

Consent for publication

All authors give consent to the publication.

Availability of data and materials

The datasets used during the current study are available from the TCGA website (https://portal.gdc.cancer.gov/).

Competing Interests

The authors declare no competing financial interests.

Funding

This work was supported by the National Natural Science Foundation of China (61471078), Dalian Science and Technology Innovation Fund (2020JJ27SN066) and the Fundamental Research Funds for the Central Universities (3132014306,3132015213, 3132017075).

Authors' contributions

Supervision: BSM and FJS. Study design: BSM, FJS and JQP. Data Collection: JQP, TX and JCL. Software & analysis: JQP, CYL and XYH. Writing-original draft: JQP. Editing: BSM, FJS, and JQP. Writing-review: All authors.

Acknowledgements

Not applicable

References

1. Britt KL, Cuzick J, Phillips K-A. Key steps for effective breast cancer prevention. Nature Reviews Cancer. 2020;20(8):417-436.
2. Wild C, Weiderpass E, Stewart B. World cancer report: cancer research for cancer prevention. Lyon: International Agency for Research on Cancer. 2020:23-33.
3. Thompson D, Easton D. The genetic epidemiology of breast cancer genes. Journal of mammary gland biology and neoplasia. 2004;9(3):221-236.
4. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nature genetics. 2018;50(7):968-978.
5. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. Journal of the National Cancer Institute. 2008;100(14):1037-1041.
6. Maas P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, 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.

7. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of breast cancer risk based on profiling with common genetic variants. JNCI: Journal of the National Cancer Institute. 2015;107(5).
8. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature genetics. 2018;50(9):1219-1224.
9. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. The American Journal of Human Genetics. 2019;104(1):21-34.
10. Dor Y, Cedar H. Principles of DNA methylation and their implications for biology and medicine. The Lancet. 2018;392(10149):777-786.
11. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS computational biology. 2017;13(5):e1005457.
12. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiology and Prevention Biomarkers. 2005;14(8):1847-1850.
13. Alegría-Torres JA, Baccarelli A, Bollati V. Epigenetics and lifestyle. Epigenomics. 2011;3(3):267-277.
14. Wild CP. The exposome: from concept to utility. International journal of epidemiology. 2012;41(1):24-32.
15. Sun YV, Hu Y-J. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. Advances in genetics. 2016;93:147-190.
16. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome research. 2014;24(9):1550-1557.
17. Liu J, Wang K, Ma S, Huang J. Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method. Statistics and its interface. 2013;6(1):99.
18. Lello L, Avery SG, Tellier L, Vazquez AI, de Los Campos G, Hsu SD. Accurate genomic prediction of human height. Genetics. 2018;210(2):477-497.
19. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nature Protocols. 2020;15(9):2759-2772.
20. Erenpreisa J, Giuliani A. Resolution of complex issues in genome regulation and cancer requires non-linear and network-based thermodynamics. International journal of molecular sciences. 2020;21(1):240.
21. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 2017;30:3146-3154.
22. Zhu E, Jiang F, Liu C, Xu J. Partition Independent Set and Reduction-Based Approach for Partition Coloring Problem. IEEE Transactions on Cybernetics. 2020. doi: 10.1109/TCYB.2020.3025819.
23. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology. 2015;19(1A):A68.

24. Rahimi A, Gönen M. Discriminating early-and late-stage cancers using multiple kernel learning on gene sets. Bioinformatics. 2018;34(13):i412-i421.
25. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nature biotechnology. 2014;32(7):644-652.
26. Liu B, Liu Y, Pan X, Li M, Yang S, Li SC. DNA methylation markers for pan-cancer prediction by deep learning. Genes. 2019;10(10):778.
27. Ma B, Meng F, Yan G, Yan H, Chai B, Song F. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Computers in biology and medicine. 2020;121:103761.
28. Weiss A, Chavez-MacGregor M, Lichtensztajn DY, Yi M, Tadros A, Hortobagyi GN, et al. Validation study of the American Joint Committee on Cancer eighth edition prognostic stage compared with the anatomic stage in breast cancer. JAMA oncology. 2018;4(2):203-209.
29. De'ath G, Fabricius KE. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology. 2000;81(11):3178-3192.
30. Pan J. The python program of PRS based on multiple omics data and LightGBM. https://github.com/lab319/PRS_BRCA_omics_LightGBM Accessed 11 Feb 2021.
31. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. The Annals of statistics. 2010;38(2):894-942.
32. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-288.
33. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology). 2005;67(2):301-320.
34. Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and computing. 2004;14(3):199-222.
35. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. The american journal of human genetics. 2015;97(4):576-592.
36. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genetic epidemiology. 2017;41(6):469-480.
37. Snoek J, Larochelle H, Adams RP, editors. Practical Bayesian optimization of machine learning algorithms. 26th Annual Conference on Neural Information Processing Systems 2012, NIPS 2012; 2012.2951-2959
38. Alves A. Stacking machine learning classifiers to identify Higgs bosons at the LHC. Journal of Instrumentation. 2017;12(05):T05005.
39. Pavlyshenko B, editor Using stacking approaches for machine learning models. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP); 2018: IEEE.255-258
40. Barendregt JJ, Doi SA, Lee YY, Norman RE, Vos T. Meta-analysis of prevalence. J Epidemiol Community Health. 2013;67(11):974-978.

41. Rich JT, Neely JG, Paniello RC, Voelker CC, Nussenbaum B, Wang EW. A practical guide to understanding Kaplan-Meier curves. Otolaryngology—Head and Neck Surgery. 2010;143(3):331-336.
42. Polikar R. Ensemble learning. Ensemble machine learning: Springer; 2012. p. 1-34.
43. Pan Y, Liu G, Zhou F, Su B, Li Y. DNA methylation profiles in cancer diagnosis and therapeutics. Clinical and experimental medicine. 2018;18(1):1-14.
44. Cheng Y, He C, Wang M, Ma X, Mo F, Yang S, et al. Targeting epigenetic regulators for cancer therapy: Mechanisms and advances in clinical trials. Signal transduction and targeted therapy. 2019;4(1):1-39.
45. Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: Computational challenges and opportunities. Experimental & Molecular Medicine. 2020;52(9):1452-1465.
46. Rodon J, Soria J-C, Berger R, Miller WH, Rubin E, Kugel A, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nature medicine. 2019;25(5):751-758.

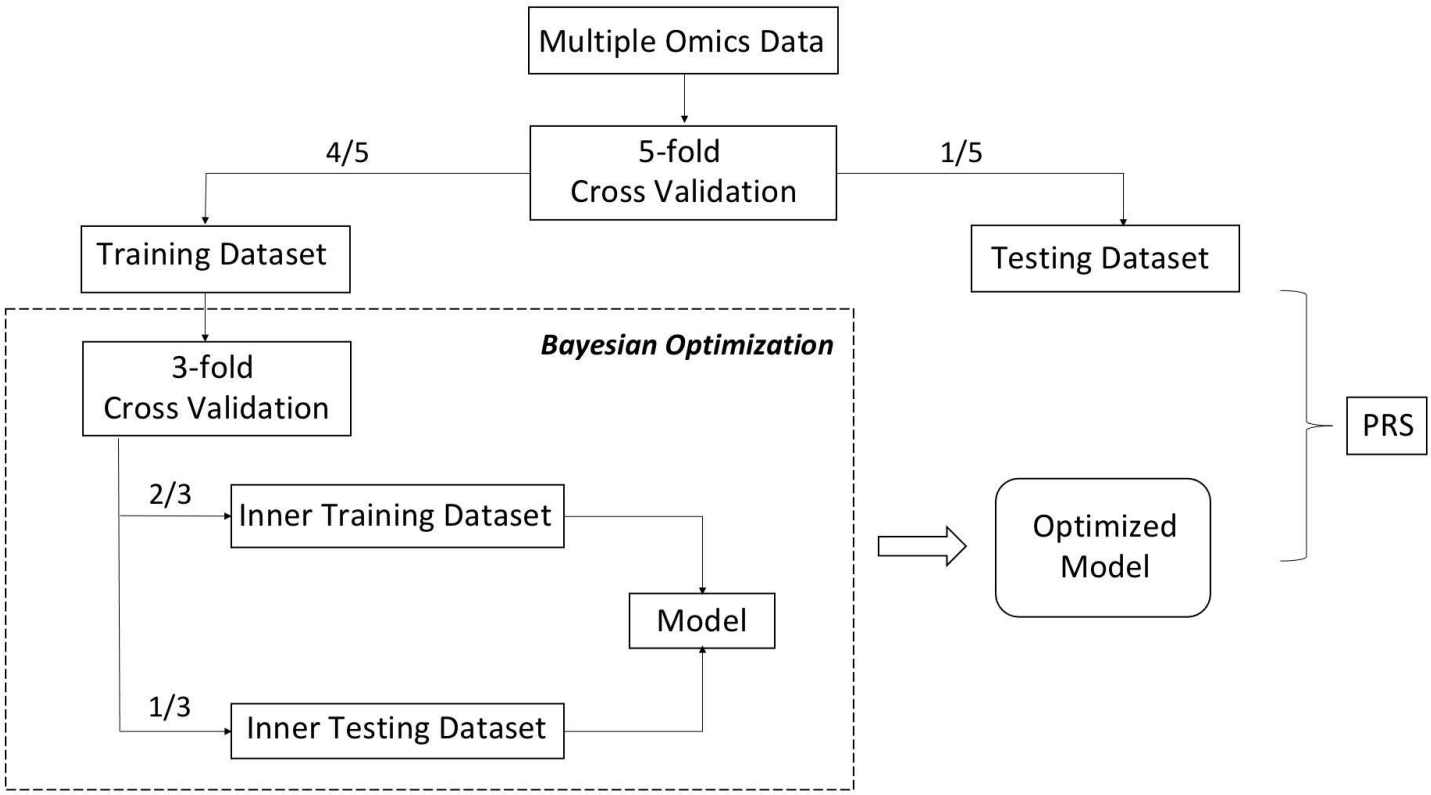
Figure Legends

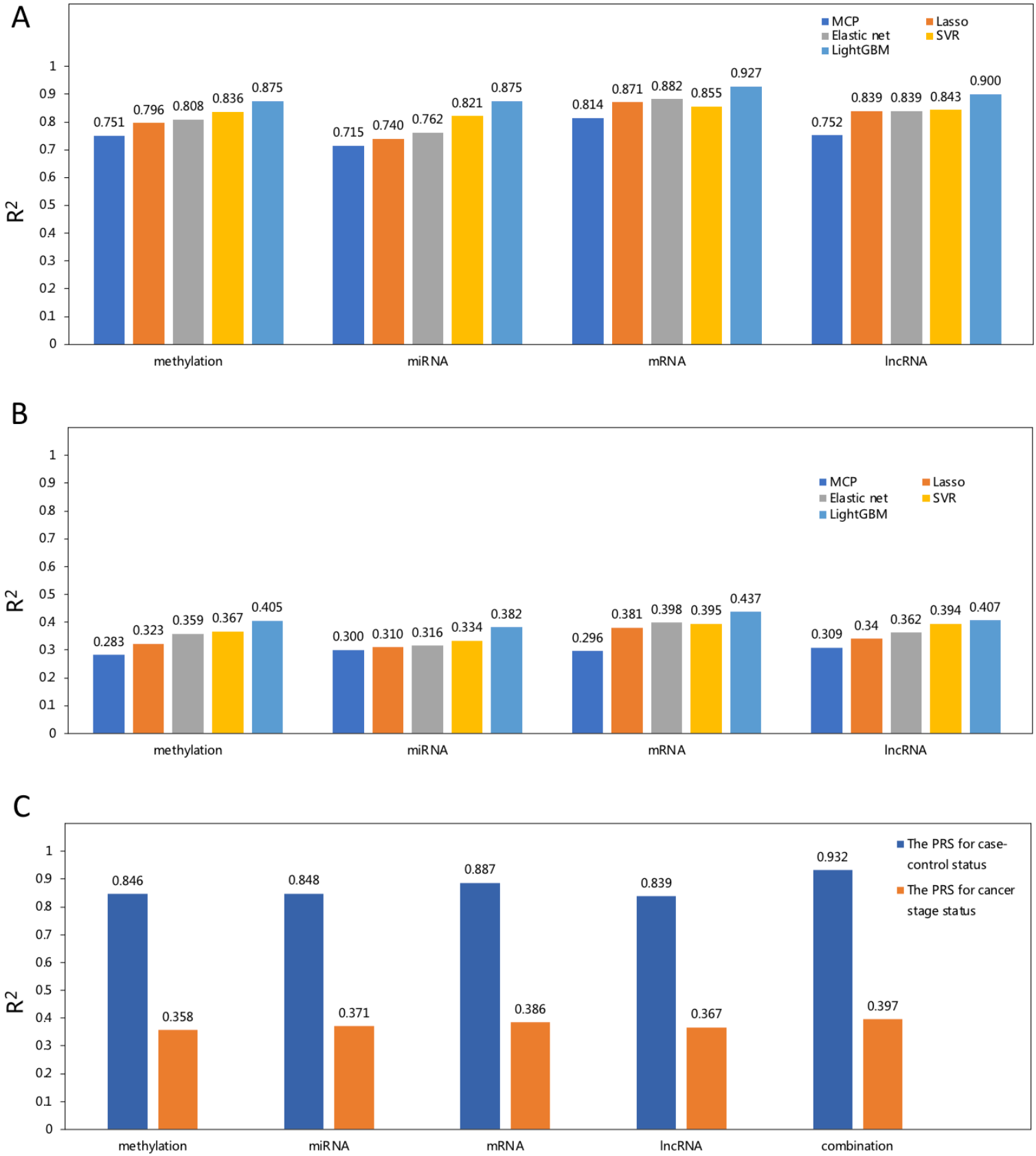
Figure 1. Schematic overview of the framework for constructing PRS model based on multiple omics data. The dataset of BRCA was split into two groups as training dataset and testing dataset based on 5-fold cross validation. We constructed PRS model by using MCP, LASSO, elastic net, SVM and LightGBM based on training dataset. The hyper-parameters of five models were optimized by using bayesian optimization and 3-fold cross validation. The PRS of testing dataset was predicted by optimized model. The predictive performance of final models was evaluated with R^2.

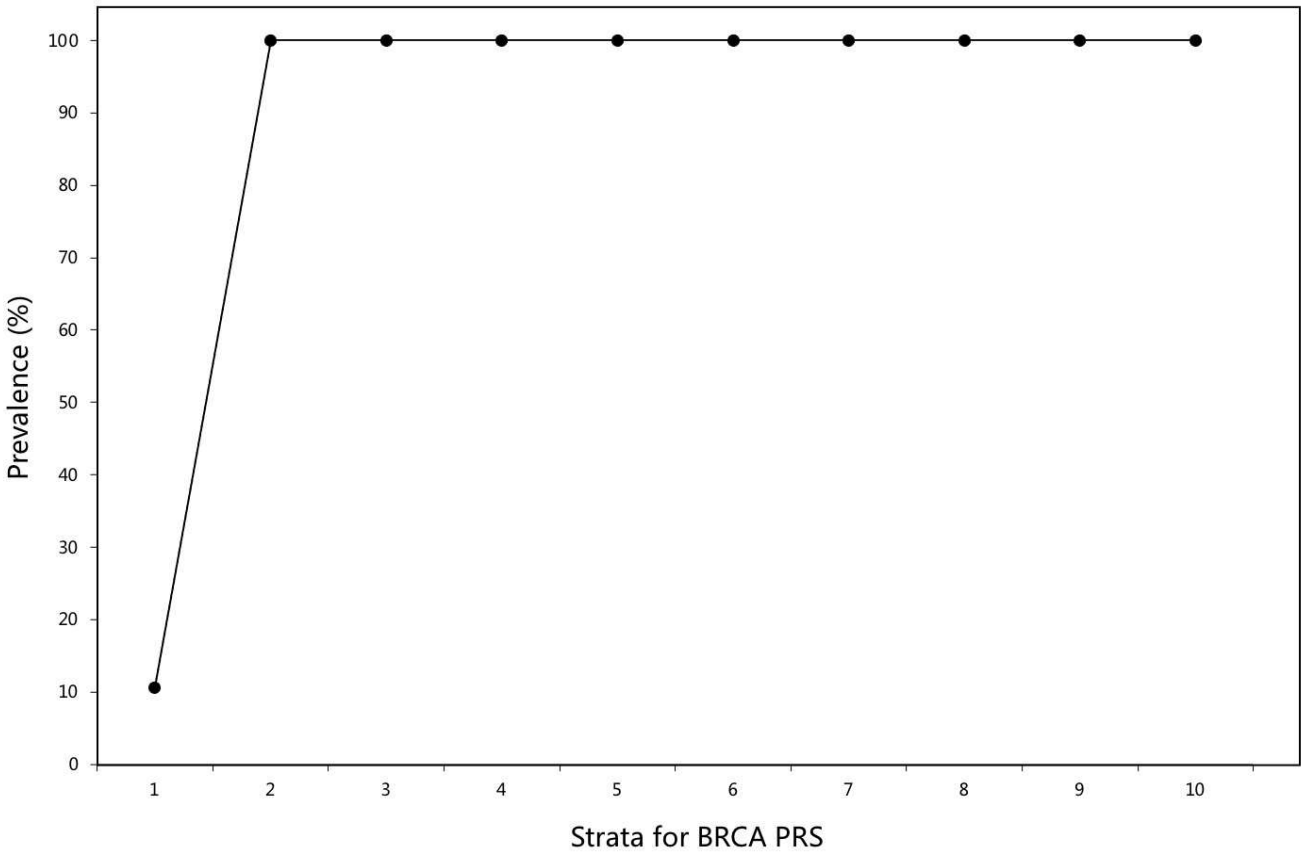
Figure 2. Predictive performance of MCP, LASSO, elastic net, SVR and LightGBM in four kinds of omics datasets. (A) Comparison results of multiple omics datasets for case-control status. (B) Comparison results of multiple omics datasets for cancer stage status. (C) Comparison results of multiple omics datasets and combination model in the common samples for case-control and cancer stage status.

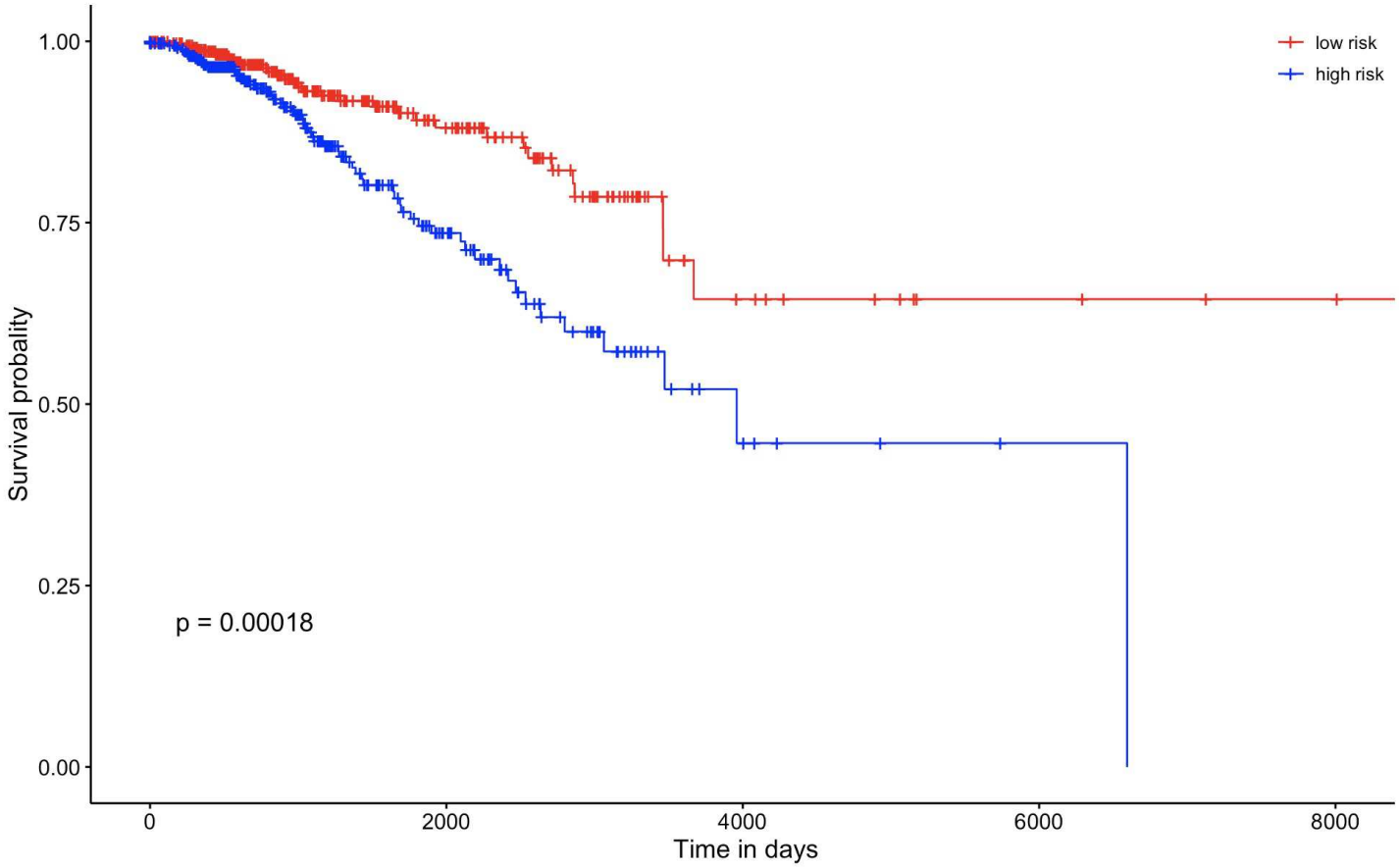
Figure 3. Prevalence strata plot of increasing PRS for case-control status. The sample size of 10 strata was equal and the prevalence of BRCA increased with the increase of PRS. The 1st stratum can be regarded as a low-risk PRS stratum and the 2nd to 10th stratum as a high-risk stratum.

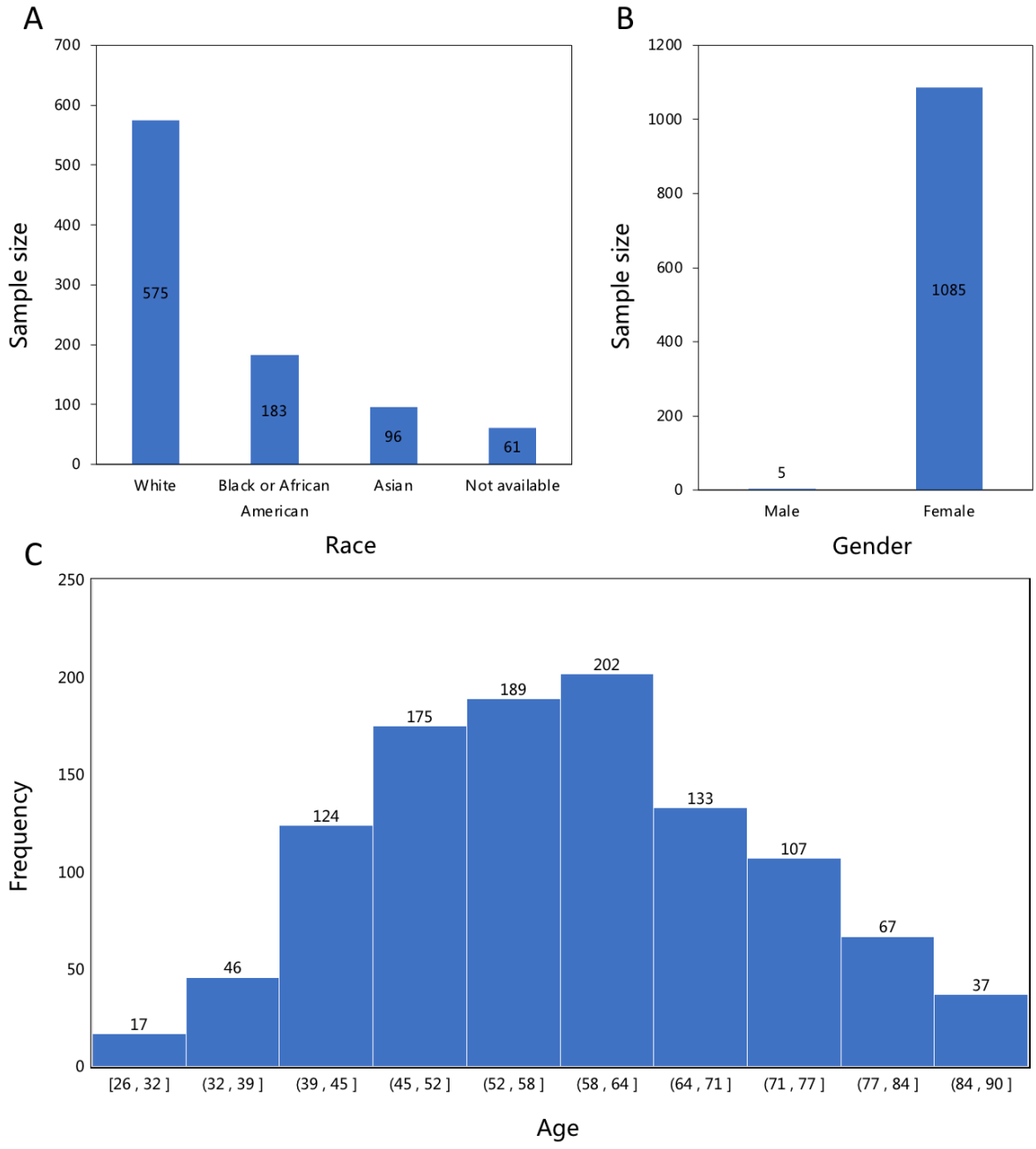
Figure 4. The KM survival curve of BRCA patients in the high-risk and low-risk groups. We divided patients into high-risk and low-risk groups based on the 50th PRS. The patients with low-risk group have better prognosis than those with high-risk group.


Multiple Omics Data
5-fold
Cross Validation
4/5
1/5
Training Dataset
Testing Dataset
Bayesian Optimization
3-fold
Cross Validation
2/3
Inner Training Dataset
Model
1/3
Inner Testing Dataset
Optimized
Model
PRS

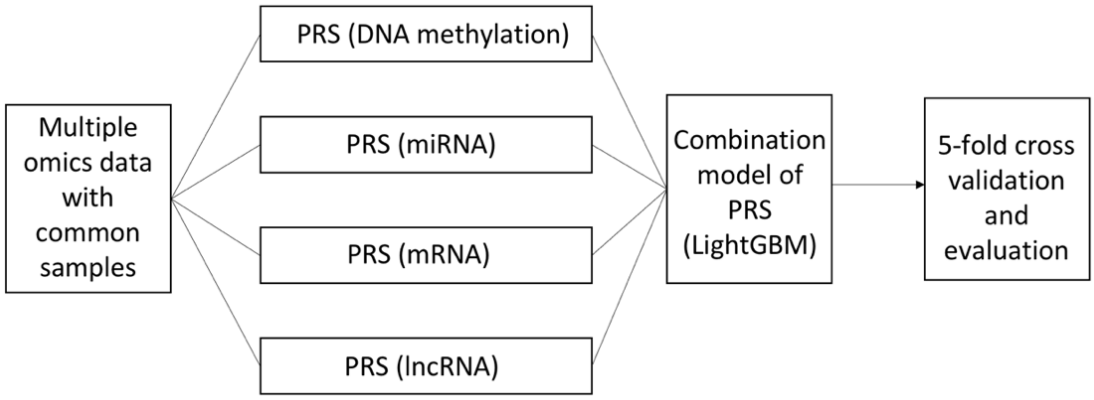

A
MCP
Lasso
Elastic net
SVR
LightGBM
R2
0.751
0.796
0.808
0.836
0.875
0.715
0.740
0.762
0.821
0.875
0.814
0.871
0.882
0.855
0.927
0.752
0.839
0.839
0.843
0.900
methylation
miRNA
mRNA
lncRNA
B
MCP
Lasso
Elastic net
SVR
LightGBM
R2
0.283
0.323
0.359
0.367
0.405
0.300
0.310
0.316
0.334
0.382
0.296
0.381
0.398
0.395
0.437
0.309
0.34
0.362
0.394
0.407
methylation
miRNA
mRNA
lncRNA
C
The PRS for case-control status
The PRS for cancer stage status
R2
0.846
0.358
0.848
0.371
0.887
0.386
0.839
0.367
0.932
0.397
methylation
miRNA
mRNA
lncRNA
combination


Prevalence (%)
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Strata for BRCA PRS


low risk
high risk
1.00
0.75
0.50
0.25
0.00
Survival probality
p = 0.00018
0
2000
4000
6000
8000
Time in days



Supplementary Figure 1. Statistics of basic demographic characteristics of patients in BRCA datasets. (A) Sample size of race including white, Black or African American, Asian and not available. (B) Sample size of



Supplementary Figure 2. Schematic overview of the framework for constructing combination model of PRS. We utilized proposed PRS model based on each omic data (DNA methylation, miRNA, mRNA and lncRNA) as new biological variables. The combination model was constructed by LightGBM model. The hyper-parameters of combination model were optimized by using bayesian optimization and 3-fold cross validation. The predictive performance was evaluated by 5-fold cross validation.

Figures

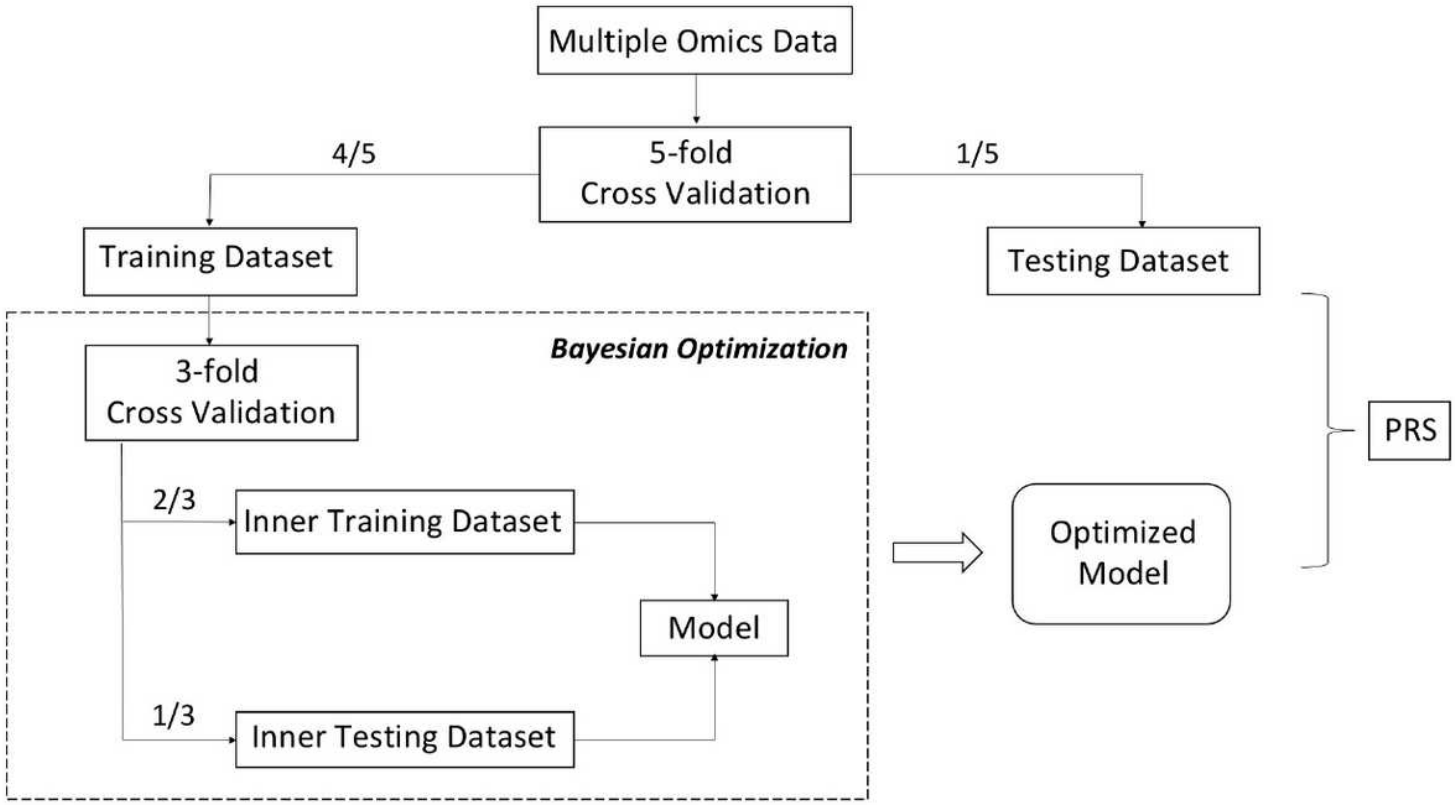


Figure 1

Schematic overview of the framework for constructing PRS model based on multiple omics data. The dataset of BRCA was split into two groups as training dataset and testing dataset based on 5-fold cross validation. We constructed PRS model by using MCP, LASSO, elastic net, SVM and LightGBM based on training dataset. The hyper-parameters of five models were optimized by using bayesian optimization and 3-fold cross validation. The PRS of testing dataset was predicted by optimized model. The predictive performance of final models was evaluated with R2.

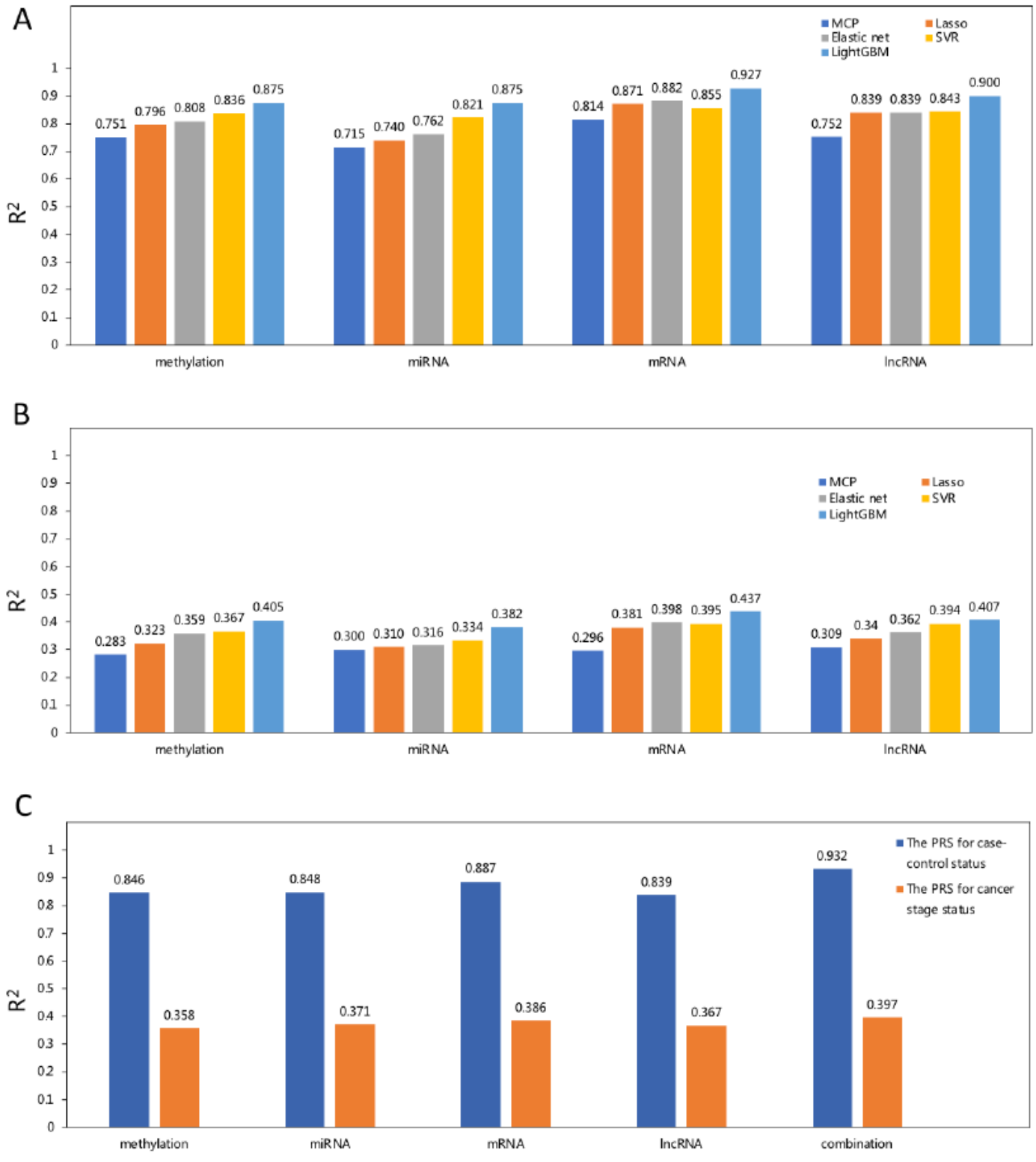


Figure 2

Predictive performance of MCP, LASSO, elastic net, SVR and LightGBM in four kinds of omics datasets. (A) Comparison results of multiple omics datasets for case-control status. (B) Comparison results of multiple omics datasets for cancer stage status. (C) Comparison results of multiple omics datasets and combination model in the common samples for case-control and cancer stage status.

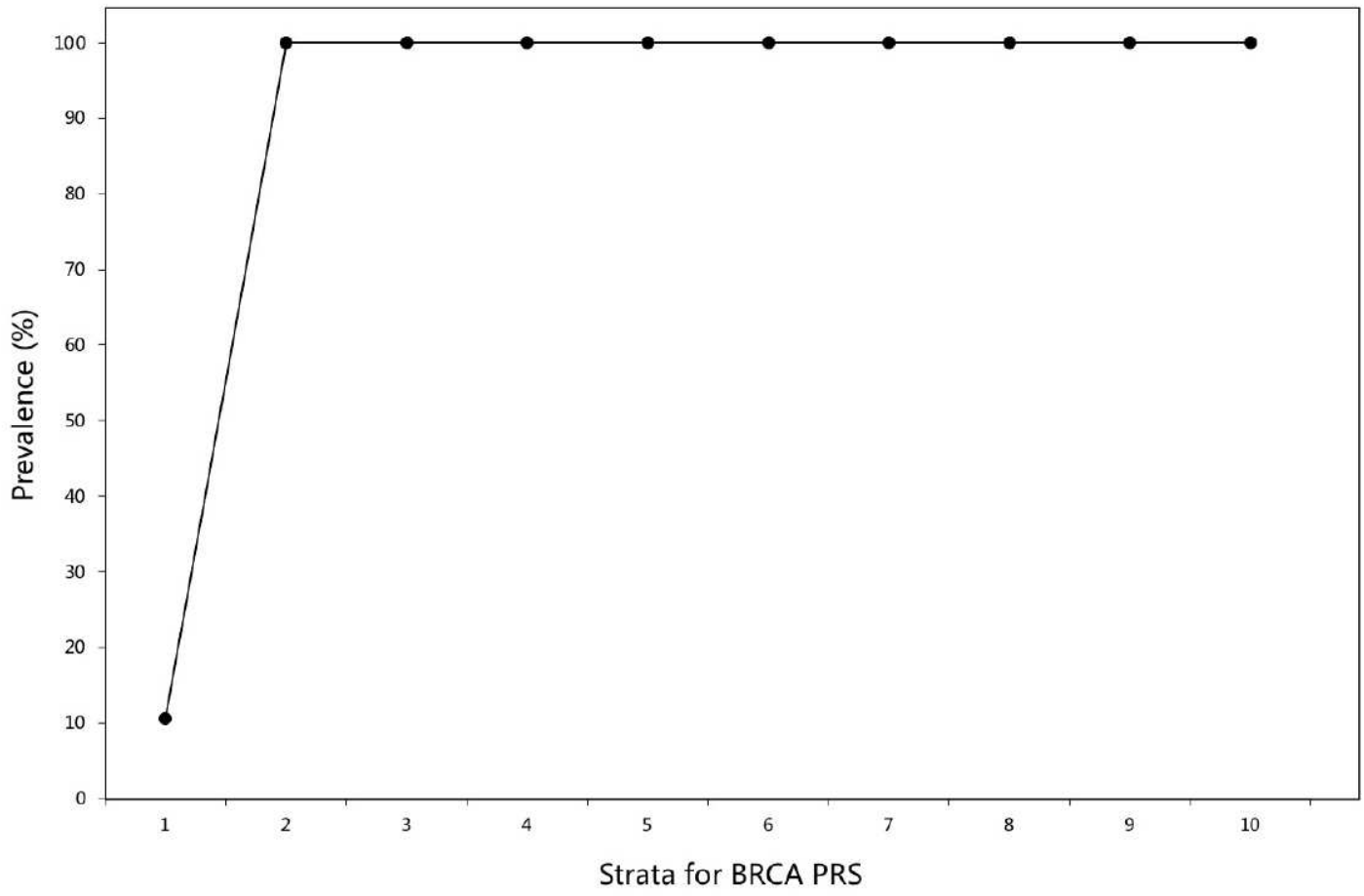


Figure 3

Prevalence strata plot of increasing PRS for case-control status. The sample size of 10 strata was equal and the prevalence of BRCA increased with the increase of PRS. The 1st stratum can be regarded as a low-risk PRS stratum and the 2nd to 10th stratum as a high-risk stratum.

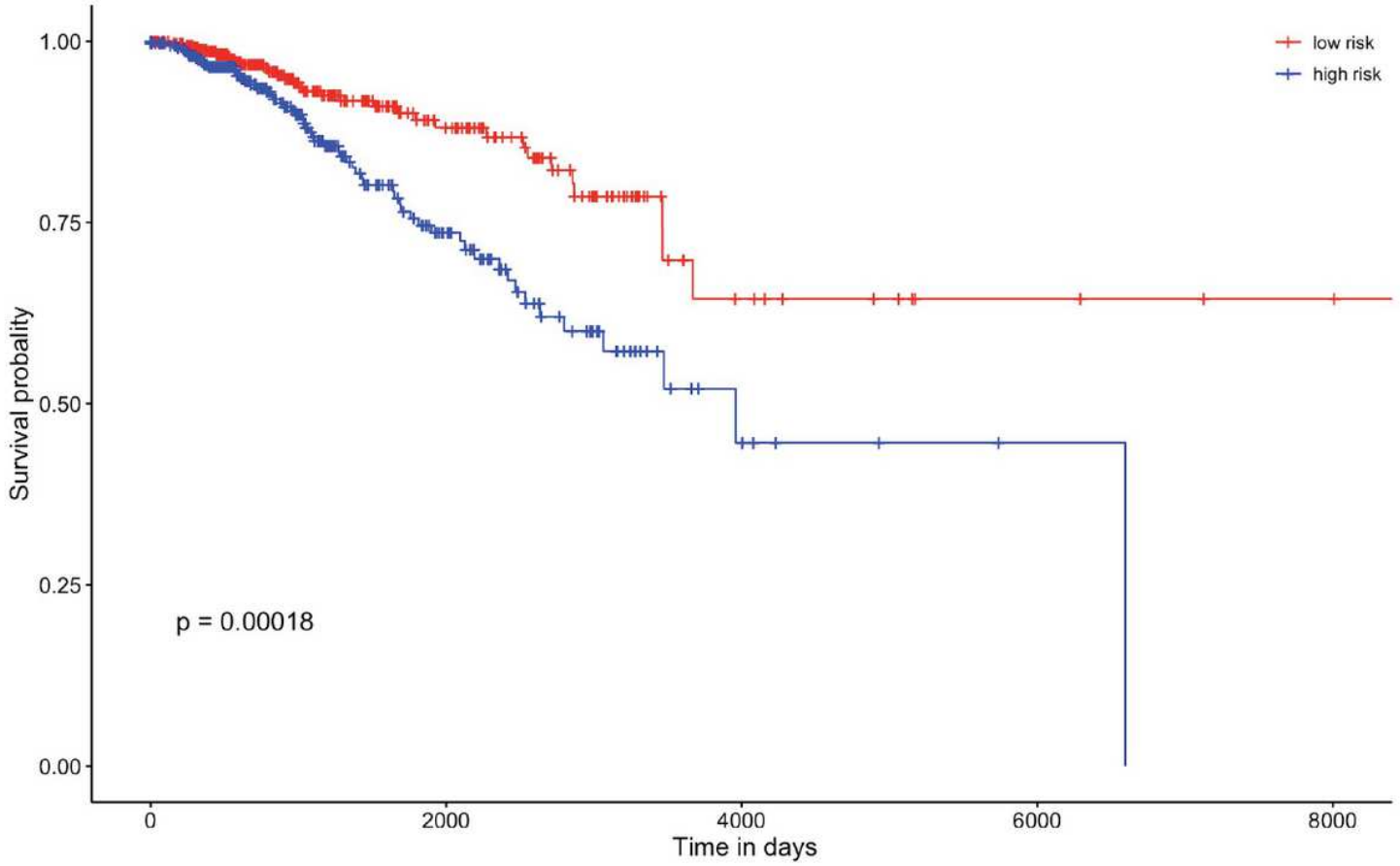


Figure 4

The KM survival curve of BRCA patients in the high-risk and low-risk groups. We divided patients into high-risk and low-risk groups based on the 50th PRS. The patients with low-risk group have better prognosis than those with high-risk group.