Predicting Future Mental Disorders Based on Plasma Proteins and Polygenic Risk Score*

Wang Jie[1], Li Yihan[1], Abudunaibi Wupuer[2], Peng Xing[2], Zhao Jianping[1], Yang Lei[2†]

(1. School of Mathematics and System Science, Xinjiang University, Urumqi Xinjiang 830017, China;
2. School of Public Health, Xinjiang Medical University, Urumqi Xinjiang 830017, China)

Abstract: Traditional psychiatric diagnosis relies on subjective symptom assessment, lacking objective biomarkers that hinder early detection and personalized treatment. Plasma proteins and polygenic risk score (PRS), as potential predictive tools, hold promise for advancing early diagnosis of mental disorders. This study aims to evaluate the predictive potential of proteomic features and PRS in multiple mental illnesses (depression, schizophrenia, and post-traumatic stress disorder (PTSD)). Using participant data from the UK Biobank-Pharma Proteomics Project, we screen protein associations with mental disorders through least absolute shrinkage and selection operator (LASSO) analysis and construct a Cox regression risk prediction model by integrating the PRS. Additionally, we evaluate predictive performance using 6 machine learning methods and Kaplan-Meier survival curves. Our findings reveal distinct predictive patterns across disorders. For depression, integrating plasma proteins with PRS significantly improves prediction beyond the clinical model (C-index=0.632 2). For schizophrenia, adding plasma proteins enhances predictive performance, whereas PRS provides no significant improvement. For PTSD, neither plasma proteins nor PRS add substantial predictive value beyond clinical variables. Risk stratification analysis demonstrat that all three mental disorders models can clearly distinguish high-risk from low-risk groups (depression: HR=2.34, P<0.001; schizophrenia: HR=5.47, P<0.001; PTSD: HR=3.02, P<0.001). Although it shows good performance in short-term prediction, its long-term prediction ability has decreased, and it needs to be further optimized in the future. This study underscores the differential utility of biomarkers across mental disorders and provides a rationale for disorder-specific predictive modeling in precision psychiatry.

Key words: plasma proteomics; polygenic risk score; mental disorders; predictive model

DOI: 10.13568/j.cnki.651094.651316.2025.09.15.0001

CLC Number: R749 **Document Code**: A **Article ID**: 2096-7675(2026)01-0001-015

引文格式: 王杰,李逸晗,阿卜杜乃比·吾普尔,彭星,赵建平,杨蕾. 基于血浆蛋白和多基因风险评分预测未来精神障碍[J]. 新疆大学学报(自然科学版中英文),2026,43(1):1-15.

英文引文格式: Wang Jie, Li Yihan, Abudunaibi Wupuer, Peng Xing, Zhao Jianping, Yang Lei. Predicting future mental disorders based on plasma proteins and polygenic risk score[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English),2026,43(1):1-15.

* **Received Date**: 2025-09-15; **Revised Date**: 2026-01-05; **Accepted Date**: 2026-01-06.

Foundation Item: The National Natural Science Foundation of China-Regional Science Foundation Project "Identification of novel drug targets for lung cancer via Mendelian randomization analysis based on blood proteomics" (62362062); The 2025 Xinjiang University Excellent Graduate Innovation Project "Research on identification of therapeutic targets and predictive factors for mental disorders based on proteomics" (XJDX2025YJS151).

Biography: Wang Jie (1999—), male, master student, research fields: bioinformatics analysis, applied statistics, E-mail: 1336831741@qq.com.

† **Corresponding Author**: Yang Lei (1981—), female, associate professor, research fields: epidemiology of the elderly, causal inference, E-mail: yanglei_616@xjmu.edu.cn.

基于血浆蛋白和多基因风险评分预测未来精神障碍

王 杰[1], 李逸晗[1], 阿卜杜乃比·吾普尔[2], 彭 星[2], 赵建平[1], 杨 蕾[2]

(1. 新疆大学 数学与系统科学学院,新疆 乌鲁木齐 830017; 2. 新疆医科大学 公共卫生学院,新疆 乌鲁木齐 830017)

摘 要:传统精神疾病诊断依赖主观症状评估,缺乏客观生物标志物,这阻碍了疾病的早期发现和个性化治疗. 作为潜在预测工具的血浆蛋白和多基因风险评分(PRS),在推进精神障碍早期诊断方面展现出巨大潜力. 本文旨在评估蛋白质组学特征与PRS在多种精神疾病(抑郁症、精神分裂症及创伤后应激障碍(PTSD))中的预测价值. 基于英国生物样本库药物基因组学项目(UK Biobank-Pharma Proteomics Project)的参与者数据,通过最小绝对值收敛和选择算子(LASSO)分析筛选出与精神障碍相关的蛋白质,并整合多基因风险评分构建Cox回归风险预测模型. 此外,采用6种机器学习方法和Kaplan-Meier生存曲线评估了模型的预测性能. 研究发现不同疾病存在显著差异:抑郁症患者中,血浆蛋白与PRS的联合应用显著提升了临床模型(C-index=0.632 2)的预测效能;精神分裂症患者中,血浆蛋白的加入虽能增强预测效果,但PRS未带来显著提升;而PTSD患者中,无论是血浆蛋白还是PRS,均未在临床变量基础上产生实质性预测价值. 风险分层分析显示,这3种精神障碍模型均能显著区分高危与低危人群(抑郁症:*HR*=2.34,*P*<0.001;精神分裂症:*HR*=5.47,*P*<0.001;PTSD:*HR*=3.02,*P*<0.001). 尽管该模型在短期预测中表现优异,但其长期预测能力有所下降,未来需要进一步优化改进. 本文揭示了不同精神障碍中生物标志物的差异性效用,并为精准精神病学中针对特定障碍的预测建模提供了理论依据.

关键词:血浆蛋白质组学;多基因风险评分;精神障碍;预测模型

0 Introduction

Mental disorders rank among the leading causes of chronic diseases, disabilities, morbidity and mortality, posing a major global public health challenge. With approximately 1 in 8 people worldwide suffering from mental illnesses, these conditions account for 14% of the global disease burden[1]. Mental disorders is also a major cause of disability, accounting for about 5% of the world's lost life years due to disability. In 2019 alone, more than 125 million years were lost[2].

However, the traditional diagnosis of mental illness mainly relies on subjective symptom assessment, which cannot fully capture the heterogeneity and comorbidity of the disease, hinder the exploration of etiology and precision treatment, and lack the support of objective biomarkers[3]. The high prevalence, disability rates, and socioeconomic burden of mental disorders urgently require more effective early diagnosis and intervention methods. By integrating multidimensional biomarkers to build precise predictive models, this approach provides new insights for early identification, differential diagnosis, and personalized treatment of mental illnesses, potentially overcoming the subjective limitations inherent in traditional psychiatric diagnostics[4].

Plasma proteins can be objectively measured and can reflect the current health status, presenting an overall disease profile[5]. They also serve as a primary reservoir for biomarkers and therapeutic targets, offering easy access and inherent disease prediction potential[6]. Research has shown that proteins are associated with the prevalence of various complex diseases, including mental disorders[7]. However, previous proteomics studies have had limited sample sizes, making it unclear whether plasma proteomics alone or in combination can provide clinically useful predictive or mechanistic information for mental disorders[8]. Polygenic risk score (PRS) integrates thousands of individual genetic loci in the human genome, using genetic variations to estimate an individual's susceptibility to disease. It then weights these loci based on the effect sizes derived from genome-wide association studies, thereby enhancing the capabilities of genetic analysis[9]. Murray et al.[10] conducted a 20-year cohort study involving patients with first-episode psychotic disorders and healthy individuals, and found that the schizophrenia PRS in the psychotic disorder group was significantly higher than that in the healthy group. Currently, PRS is widely used to construct risk models for predicting new

or common diseases, but further research is needed to determine if PRS can improve the predictive power of risk models for mental disorders. Combining the PRS for bipolar disorder with clinical data can predict the disease risk of offspring of bipolar disorder patients with high accuracy (*AUC*=0.81)[11]. These studies all indicate the potential clinical application value of PRS in high-risk populations, providing possibilities for the early identification and intervention of high-risk individuals. However, it should be noted that the current intervention methods for high-risk populations are still relatively limited[12], and there is a lack of relevant analysis combining plasma protein data.

We utilized data from the UK Biobank to illustrate the contributions of PRS and proteomics to predicting the onset risk of 5 mental disorders, including major depressive disorder, schizophrenia, and post-traumatic stress disorder (PTSD). The study systematically evaluates the predictive value of protein levels and PRS for mental health risks. It will provide a more accurate scientific basis for early screening and personalized intervention in the spiritual field, and is expected to promote the development of psychiatry.

1 Materials and Methods

1.1 Study Design and Study Participants

1.1.1 Study Design

Firstly, we used cohort study data from the UKB-PPP (UK Biobank-Pharma Proteomics Project) to develop, validate and compare protein and non-protein predictive models. Secondly, PRS was integrated to develop a prediction model for the risk of developing mental disorders based on prospective population cohorts. Finally, the combined model was evaluated in predictive performance, predictor importance, risk stratification analysis and longitudinal analyses. A flowchart shows the overall study design (Figure 1).

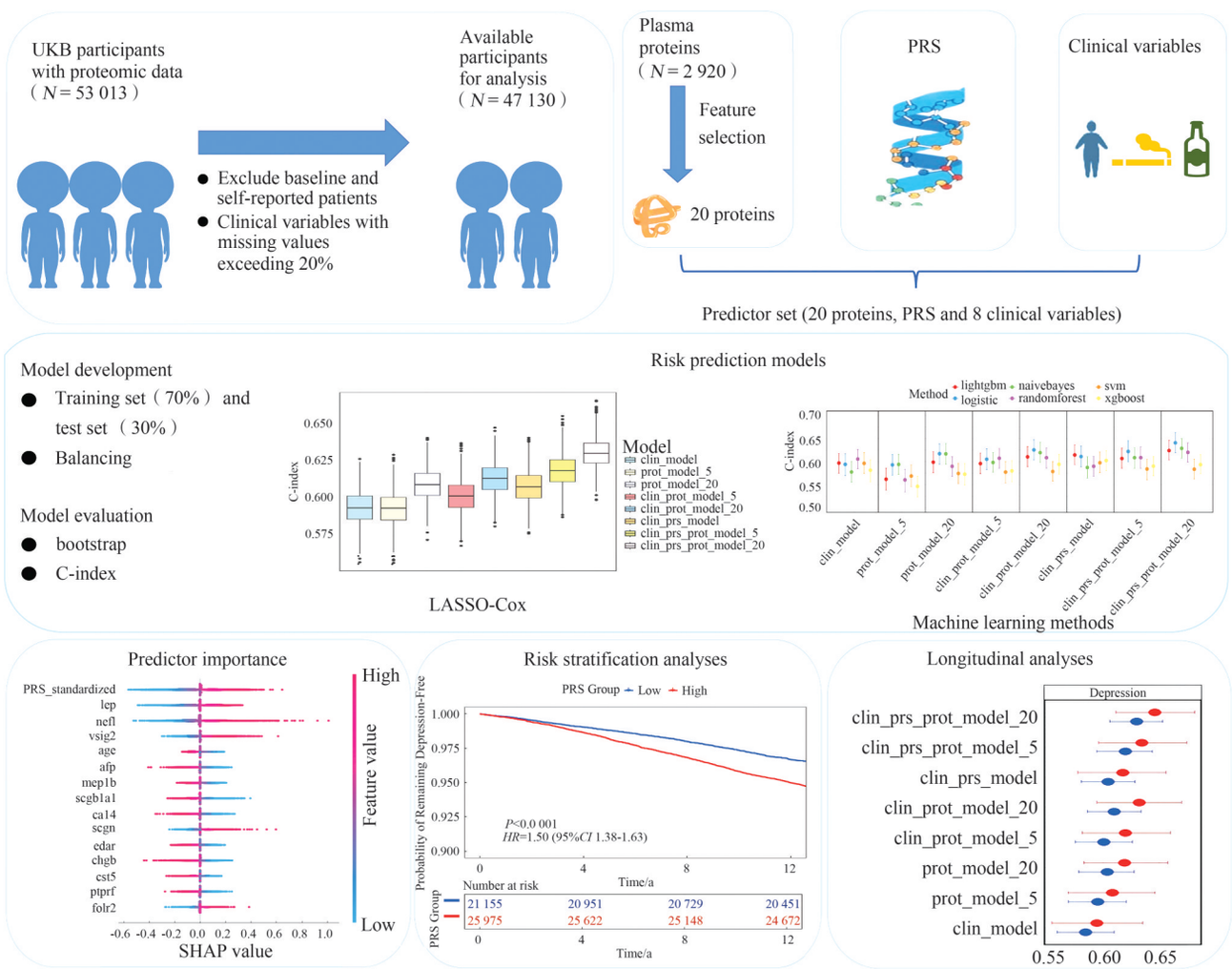

Figure 1 Study flowchart

1.1.2 Plasma Proteomics

In October 2023, UKB unveiled a critical resource from its Pharma Proteomics Project (UKB-PPP): plasma protein level data collected from 53 013 blood samples during the 2006—2010 recruitment phase[13]. The research team utilized Olink Explore 3 072 proximity extension detection technology to obtain quantitative information for 2 923 circulating proteins, with approximately 17.5% of the data being missing. We employed the missForest R software package to impute protein levels across 8 detection panels (Cardiovascular, Cardiovascular II, Inflammation, Inflammation II, Neurology, Neurology II, Oncology, and Oncology II), while adjusting for age and gender factors.

1.1.3 Phenotyping

Our primary outcome includes 5 types of mental disorders: schizophrenia (F20-F29), bipolar disorder (F30-F31), depression (F32-F33), post-traumatic stress disorder (PTSD, F43.1), and eating disorders (F50). To determine the diagnosis and corresponding dates of mental disorders, we extracted relevant data from the first occurrence record (field 131022) according to the International Classification of Diseases (ICD), 10th Edition (ICD-10). These records were sourced from multiple data sources, including primary healthcare data (field 42040), hospitalization data (field 42040, 41270, and 41271), and death registry (field 40001 and 40002). Participants who had been diagnosed with any of the 5 mental disorders before baseline time were excluded. The follow-up period began on the day they visited the assessment center and continued until December 31, 2022, death, or withdrawal from the study, whichever occurred first.

1.2 Clinical Variables

At baseline, age, sex, body mass index (BMI), triglycerides, systolic blood pressure, creatinine, smoking status and alcohol intake were collected as clinical variables for the model. The missing rate of clinical variables was less than 20%, and the missing values were interpolated using the K-Nearest Neighbor interpolation method.

1.3 Construction of PRS

The 5 mental disorders in the whole-genome summary statistics include bipolar disorder[14] and depression[15], which are from the Psychiatric Genomics Consortium (PGC, https://www.med.unc.edu/pgc/download-results); eating disorders, post-traumatic stress disorder, and schizophrenia, which are from the Finnish database. To obtain the association effect sizes of single nucleotide polymorphisms (SNPs), we used PRSice-2 to calculate the PRS for these 5 mental disorders. We calculated the PRS for each participant using the P+T method. To identify the most predictive causal variants for PRS modeling, we processed the genome-wide association studies (GWAS) data by selecting SNPs with minor allele frequencies (MAF) greater than 0.01 and variant information (INFO) greater than 0.8, and removed ambiguous SNPs.

1.4 Statistical Analyses

1.4.1 Baseline Characteristics Description

In this study, the group characteristics of participants were evaluated by chi-square test and t-test. The median and four-quarter distance were used to represent the continuous variables, and the frequency and percentage were used to present the categorical variables.

1.4.2 Development of Prediction Models

For each mental disorder, feature selection was performed on 2 920 proteins. The selected protein set was screened using the least absolute shrinkage and selection operator (LASSO) regression with 50 subsamples from the feature selection set. To achieve data-driven selection of the optimal feature set, our machine learning framework reduced the coefficients of strongly correlated variables to 0. In each iteration, we used the caret R package to perform a 5-fold cross-validation of the hyperparameter lambda by grid search. The score was calculated by summing the absolute total weight of each model with the best lambda value over 50 iterations, and was used to identify the top 20 proteins. We divided the data set into a 70% training set and a 30% validation set, and then used the

ROSE R software package to address the problem of data imbalance. Using the glmnet R package, we constructed a regularized Cox regression by combining the top 20 proteins with the highest feature selection scores, clinical variables, and PRS. We performed 5-fold cross-validation on the training set and validation on the test set. We calculated the C-index using 1 000 bootstrap samples. Similarly, we used the same method to build a regularized Cox regression model for each combination of single clinical variable, single top 5 proteins, single top 20 proteins, clinical variable plus top 5 proteins, clinical variable plus top 20 proteins, clinical variable plus PRS, and clinical variable plus top 20 proteins plus PRS. In order to compensate for the limitations of a single method, the accuracy and robustness of the model were enhanced, we also used 6 machine learning methods, including logistic regression, extreme gradient boosting (XGBoost), random forest, Naive Bayes, support vector machine (SVM) and lightGBM. The same data processing and feature combination as the regularized Cox regression model were adopted to build the model.

1.4.3 Predictor Importance

To interpret the output of the prediction model, we introduce the SHAP score. The SHAP scoring system uses game theory to build a feature importance evaluation system. The scoring mechanism quantifies the marginal contribution of each feature by generating feature power sets, and determines the feature importance by calculating the average absolute SHAP score of all model combinations on the test set. Through systematic analyses of SHAP values, we identified key factors influencing mental disorder risk and their relative importance in predictive outcomes. This approach deepened our understanding of data characteristics and their underlying correlations. Finally, we visually presented the data distribution using a swarm plot for intuitive visualization.

1.4.4 Hazards Ratios Estimate and Risk Stratification Analyses

The Kaplan-Meier curve and Cox proportional risk model were used to calculate the probability of non-morbidity over time in different groups[16]. Based on the survival ROC package[17–18], the optimal cutoff was determined for the overall analyses to accurately divide the high-risk and low-risk groups. Participants were divided into high-risk group (risk score above threshold) and low-risk group (risk score below threshold). Similarly, we used the same method to divide participants into high PRS group and low PRS group, calculate the survival probability of high PRS group and low PRS group over time, and evaluate the difference between survival curves of high PRS group and low PRS group samples by log-rank test.

1.4.5 Longitudinal Analyses

Given that an individual's protein levels are dynamic while genotypes remain static, the performance of mental disorders risk models was evaluated in both short-term (5 years) and long-term (10 years) assessments following blood sampling. In these studies, only individuals diagnosed with the disorder within the specified timeframe (5 or 10 years) were classified as event cases, while others were designated as controls.

1.4.6 Correlation Analyses and Enrichment Analyses

To ensure the robustness of the predictive model and to explore the underlying biology, 2 subsequent analyses were conducted on the selected plasma proteins. Firstly, a correlation analysis was performed to assess potential multicollinearity among the protein features. Secondly, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were carried out to systematically investigate the biological processes and pathways represented by the protein set.

2 Results

2.1 Participants' Characteristics

This study ultimately enrolled 47 130 participants from the UKB who were free of depression, schizophrenia, bipolar disorder, PTSD, or eating disorders at baseline. Among them, 24 846 (52.72%) were female. The median age at baseline assessment was 58 years (interquartile range [IQR] 50-64). During the 13.95-year follow-up period

(IQR 13.3-14.6), 2 291 cases of depression, 102 cases of schizophrenia, 380 cases of PTSD, 48 cases of bipolar disorder, and 10 cases of eating disorders were identified. The number of bipolar disorder and eating disorder cases did not meet study requirements (the data was divided into test-set cases in a 7∶3 ratio, with over 20 cases in the test set). Table 1 summarizes the baseline characteristic data of participants.

Table 1 Baseline characteristics of UKB participants included in the study

Participants characteristics	All participants	Schizophrenia		Depression		PTSD		Bipolar disorder		Eating disorders	
	N=47 130	*N*=102	*P*	*N*=2 291	*P*	*N*=380	*P*	*N*=48	*P*	*N*=10	*P*
Age [IQR]	58 [50-64]	62 [53-66]	0.008	58 [50-65]	0.912	52 [46-60]	<0.001	55 [49-64]	0.448	57 [47-65]	0.711
Sex (female)	24 846 (52.72)	46 (45.10)	0.149	1 374 (59.97)	<0.001	237 (62.37)	<0.001	26 (54.17)	0.955	8 (80.00)	0.158
Smoking status	24 252 (51.46)	58 (56.86)	0.320	1 334 (58.23)	<0.001	207 (54.47)	0.259	29 (60.42)	0.272	6 (60.00)	0.823
Alcohol intake	1.71 [0.00-3.43]	1.14 [0.00-3.43]	0.472	1.14 [0.00-3.14]	0.001	1.14 [0.00-3.43]	0.451	1.43 [0.00-3.29]	0.455	1.43 [0.57-2.29]	0.351
Body mass index	26.73 [24.16-29.79]	27.25 [24.22-31.13]	0.220	27.80 [24.86-31.44]	<0.001	27.25 [24.74-30.43]	0.026	26.81 [23.65-29.09]	0.700	26.33 [17.14-26.89]	0.006
Creatinine/ (μmol/L)	70.90 [62.00-81.66]	73.00 [61.60-83.10]	0.624	69.00 [61.10-80.20]	0.002	68.75 [60.70-79.60]	0.047	69.85 [59.25-82.15]	0.380	70.50 [61.40-83.70]	0.859
Triglyceride/ (μmol/L)	1.49 [1.06-2.11]	1.51 [1.07-2.44]	0.018	1.58 [1.13-2.23]	<0.001	1.50 [1.05-2.09]	0.265	1.43 [1.00-1.84]	0.191	0.82 [0.72-1.15]	0.023

Note: continuous data are described as median (interquartile range), and categorical variables are presented as *n* (%). The differences between the incident mental disorders group and the control group were compared using t-test for continuous variables and χ^2 tests for discrete variables. The *P* indicates the significance of the difference in clinical variables between the group with mental illness and the control group

2.2 Development and Evaluation of Prediction Model

Among the 3 mental disorders, depression and schizophrenia were each found to be associated with 20 proteins, while only 12 risk-related proteins were ultimately identified for PTSD. In the 8 models constructed for depression (Figure 2(a)), the performance of models using clinical variables alone and those using 5 proteins alone was approximately 0.594 1 (95%*CI* 0.571 8-0.616 6) and 0.594 3 (95%*CI* 0.572 3-0.617 2), respectively. When incorporating more proteins, model performance improved significantly (C-index: 0.610 6 (95%*CI* 0.589 1-0.632 0)). By comparison, combining proteins with clinical variables showed only a modest improvement in C-index (clin_prot_model_5: 0.602 5 (95%*CI* 0.579 6-0.625 0) and clin_prot_model_20: 0.614 5 (95%*CI* 0.592 4-0.635 3)). Similarly, integrating clinical variables, PRS, and proteins further enhanced model performance. The PRS component demonstrated particularly significant improvement in depression risk modeling, with all models showing that clin_prot_model_20 achieved the best results at C-index: 0.632 2 (95%*CI* 0.610 0-0.652 6). However, the models constructed for schizophrenia and PTSD did not yield similar results. In the schizophrenia risk model (Figure 2(b)), the prot_model_5 model performed worse than the clin_model. Although the C-index in the prot_model_20 model was 0.603 2 (95%*CI* 0.511 3-0.690 6), when proteins were combined with clinical variables, the performance of models containing 5 or 20 proteins did not improve. This indicates that there is information overlap between proteins and clinical variables. When comparing models incorporating PRS, it was found that PRS did not improve model performance. In the PTSD risk model (Figure 2(c)), both clin_model (C-index: 0.629 8 (95%*CI* 0.575 3-0.682 7)) and clin_prs_model

(C-index: 0.629 9 (95%*CI* 0.575 4-0.682 7)) demonstrated superior performance compared to other models, while PRS of PTSD and proteins showed weaker predictive capabilities.

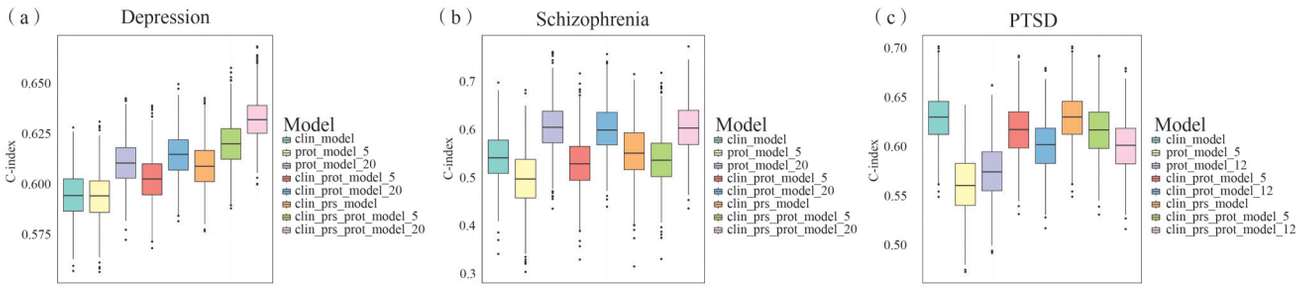

Figure 2 Cox proportional hazards regression analyses

Note: (a) C-index distribution of 1 000 bootstrap for 8 depression risk models. (b) C-index distribution of 1 000 bootstrap for 8 schizophrenia risk models. (c) C-index distribution of 1 000 bootstrap for 8 PTSD risk models. Clin_model refers to a model that only includes clinical variables. Prot_model_5 and prot_model_20 refer to the top 5 or 20 proteins in the model that are only included in feature selection. Clin_prot_model_5 and clin_prot_model_20 respectively refer to the models that incorporate clinical variables and the top 5 or top 20 protein in selection features. Clin_prs_model refers to a model that incorporates clinical variables and PRS. Clin_prs_prot_model_5 and clin_prs_prot_model_20 respectively refer to the models that incorporate clinical variables, PRS, and the top 5 or top 20 protein in selection features

In the machine learning risk model for depression, all models exhibited C-index between 0.55 and 0.65, indicating average discriminative power. Performance variations among different feature combinations and machine learning methods were relatively small, with most results showing overlapping confidence intervals. Logistic regression and Naive Bayes demonstrated superior performance in most cases, while SVM and XGBoost performed comparatively poorly (Figure 3(a)). When incorporating proteins into the models of logistic regression and Naive Bayes, their performance surpassed single-clinical-variable models. Further inclusion of PRS significantly enhanced model performance. The combined model incorporating clinical variables, PRS, and 20 proteins, achieved relatively high C-index across most machine learning methods: naivebayes-C-index: 0.628 7 (95%*CI* 0.607 5-0.649 5); randomforest-C-index: 0.620 0 (95%*CI* 0.598 3-0.641 1); logistic-C-index: 0.640 7 (95%*CI* 0.619 4-0.663 3); lightgbm-C-index: 0.623 9 (95%*CI* 0.603 8-0.646 0). In the schizophrenia machine learning risk model, the models exhibited poor performance with C-index ranging from approximately 0.28 to 0.68, indicating a wider confidence interval (Figure 3(b)). Most models performed similarly to random classification. The logistic regression model demonstrated stable and relatively better performance compared to other models, showing optimal performance among those incorporating only 20 proteins (C-index: 0.591 0 (95%*CI* 0.489 7-0.694 4)). However, the models clin_prot_model_20 and clin_prs_prot_model_20 showed C-index of 0.579 3 (95%*CI* 0.471 4-0.673 6) and 0.578 4 (95%*CI* 0.471 1-0.688 0), respectively. This indicates information overlap between proteins and clinical variables. The presence of PRS did not improve model performance, consistent with the findings of the Cox proportional hazards model. In the PTSD machine learning risk model, the C-index primarily ranged between 0.45 and 0.67, showing relatively stable results (Figure 3(c)). Logistic regression maintained its optimal performance in most models, with the clin_prs_model (C-index: 0.629 3 (95%*CI* 0.576 8-0.677 1)) demonstrating the best fit, closely matching the clin_model (C-index: 0.628 1 (95%*CI* 0.577 6-0.675 3)). Other models exhibited lower performance, with limited predictive capabilities for proteins and PRS. The evaluation of 6 machine learning algorithms shows that linear models demonstrate greater stability and consistency on the dataset compared to nonlinear models. Moreover, the prediction results of linear models are highly consistent with our primary LASSO-Cox analysis outcomes.

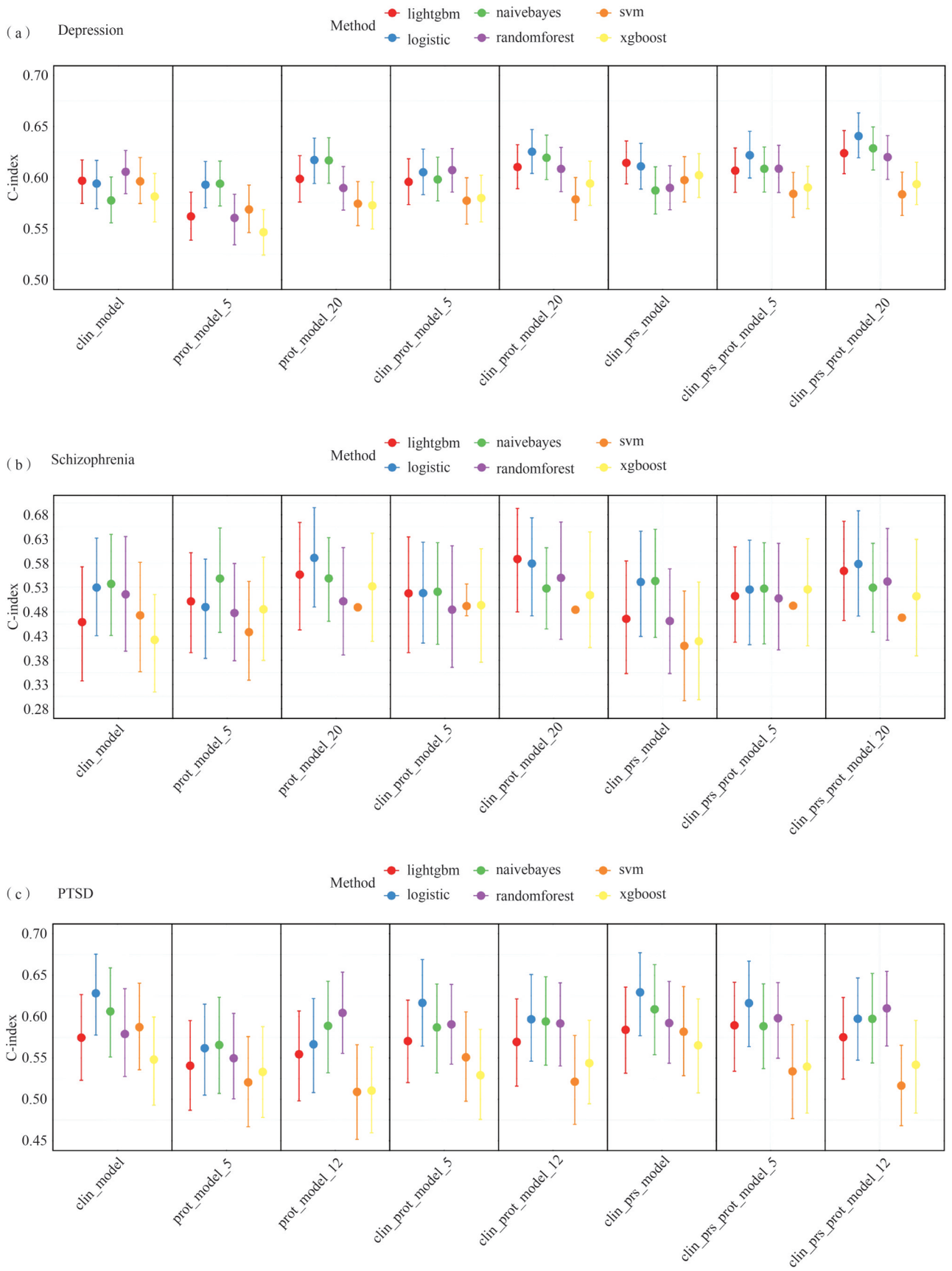

Figure 3 Performance comparison of prediction models with 8 different feature combinations under 6 machine learning methods

2.3 Predictor Importance

The SHAP diagram shows the relative importance and influence of the top 15 features included in the model clin_prs_prot_model for the 3 mental disorders. In depression risk models, PRS demonstrates the highest relative importance as a key risk factor. Elevated concentrations of proteins including leptin (LEP), neurofilament light polypeptide (NEFL), V-set and immunoglobulin domain-containing protein 2 (VSIG2), secretagogin (SCGN), and folate receptor beta (FOLR2) significantly increase depressive risk. Conversely, higher levels of alpha-fetoprotein (AFP), meprin A subunit beta (MEP1B), uteroglobin (SCGB1A1), carbonic anhydrase 14 (CA14), tumor necrosis factor receptor superfamily member EDAR (EDAR), secretogranin-1 (CHGB), cystatin-D (CST5), and receptor-type tyrosine-protein phosphatase F (PTPRF) tend to reduce depressive risk. Only age was included in the clinical variables, which was a protective factor for depression (Figure 4(a)). In the schizophrenia risk model, PRS was not included. Clinical variables such as alcohol consumption and gender were added. Protein-related factors KH domain-containing protein 3 (KHDC3L), hepatoma-derived growth factor (HDGF), killer cell immunoglobulin-like receptor 2DL3 (KIR2DL3), proline-rich protein 4 (PRR4), proline/serine-rich coiled-coil protein 1 (PSRC1), and lymphocyte antigen 96 (LY96) were identified as risk factors. The following factors were identified as protective elements: probable ATP-dependent RNA helicase (DDX4), probable G-protein coupled receptor 158 (GPR158), Interleukin-13 receptor subunit alpha-1 (IL13RA1), fatty acid-binding protein 9 (FABP9), neurocan core protein (NCAN), and cobalamin binding intrinsic factor (CBLIF). Among these, DDX4 demonstrated the highest relative significance (Figure 4(b)). In the PTSD risk model, the selected clinical variables were significantly increased, the importance of protein was relatively weak, and PRS was not included, which was consistent with the results of Cox proportional risk model (Figure 4(c)).

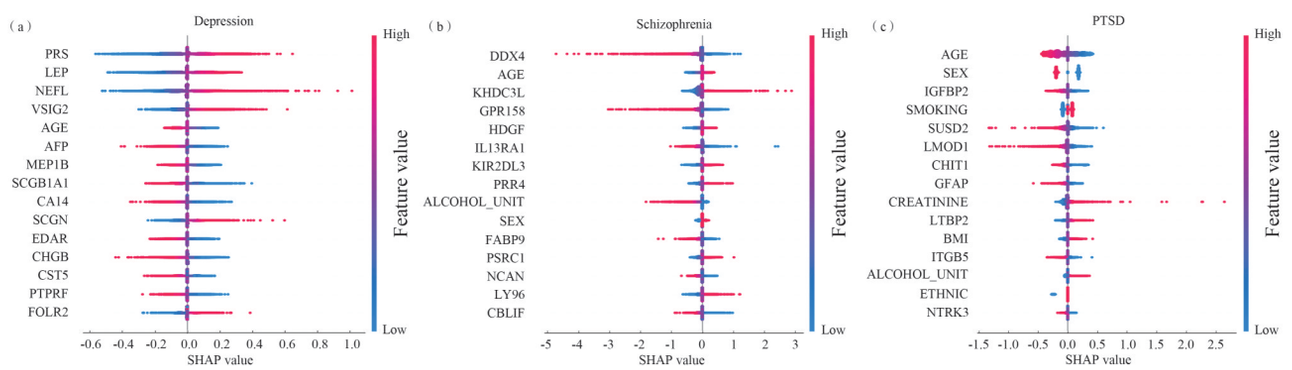

Figure 4 SHAP bee swarm plots of the three mental disorders in clin_prs_prot_model_20

Note: (a) The SHAP values of the depression in clin_prs_prot_model_20. (b) The SHAP values of the schizophrenia in clin_prs_prot_model_20. (c) The SHAP values of the PTSD in clin_prs_prot_model_12. The figure shows the SHAP results for the top 15 features of relative importance

2.4 Risk Stratification Analyses

Participants were divided into high-risk and low-risk groups according to the best critical value calculated by the largest Youden index, and the characteristics showed significant differentiation in 3 mental disorders risk models. In the depression risk model, the risk ratio (HR) of high-risk population was significantly 2.34 (95%*CI* 2.15-2.55, Figure 5(a)); In the schizophrenia risk model, the risk of high-risk population was 5.47 times that of low-risk population (95%*CI* 3.54-8.45, Figure 5(c)); In the PTSD risk model, the risk ratio of high-risk population was also 3.02 (95%*CI* 2.46-3.71, Figure 5(e)). However, when participants were divided into high and low PRS groups using the same method, only the depression risk model showed significant results, with the high PRS group demonstrating a significantly higher disease risk than the low PRS group (*HR*: 1.50 (95%*CI* 1.38-1.63), Figure 5(b)). While the stratified analyses of schizophrenia and PTSD risk models failed to pass statistical tests, all risk ratios remained greater than 1 (Figure 5(d), Figure 5(f)).

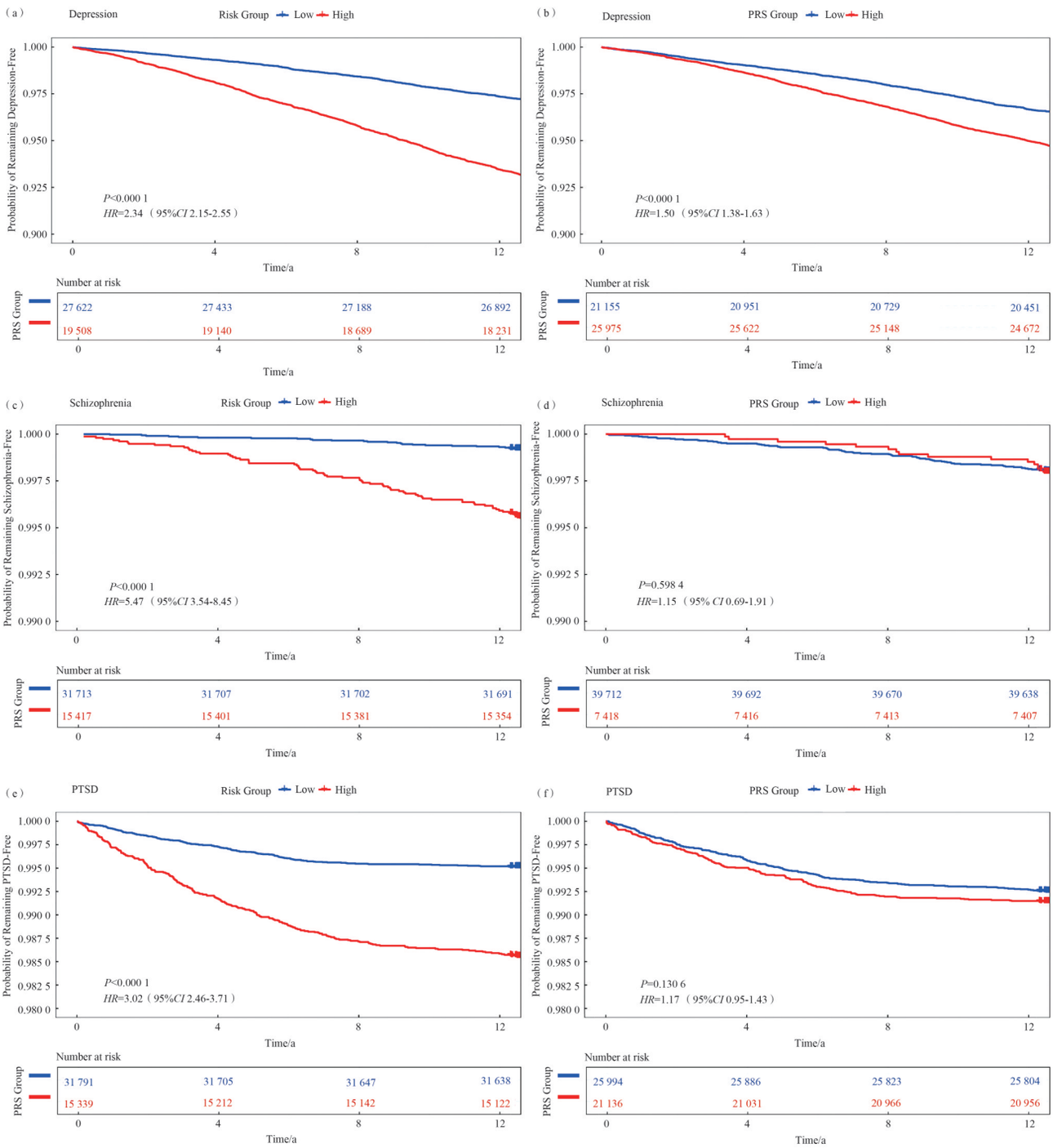

Figure 5 KM curve of risk stratification of mental disorders by risk group and PRS

Note: (a) Kaplan-Meier survival curve of depression after risk stratification based on proteome model prediction. (b) Kaplan-Meier survival curve of depression after risk stratification based on PRS model prediction. (c) Kaplan-Meier survival curve of schizophrenia after risk stratification based on proteome model prediction. (d) Kaplan-Meier survival curve of schizophrenia after risk stratification based on PRS model prediction. (e) Kaplan-Meier survival curve of PTSD after risk stratification based on proteome model prediction. (f) Kaplan-Meier survival curve of PTSD after risk stratification based on PRS model prediction

2.5 Longitudinal Analyses

We tested the models at 5 years and 10 years post blood draw (Figure 6). Among depression and PTSD cases, 8 models demonstrated gradual performance decline with extended follow-up. Most protein-containing models showed sig-

nificant performance gaps: the clin_prot_model_20 model exhibited the most marked decrease (0.631 8→0.609 9), while PTSD models also registered substantial drops (0.647 9→0.603 1), with prot_model_20 showing the sharpest decline (0.622 1→0.573 5). These findings support plasma proteomics as a biomarker reflecting biological states near onset rather than stable risk traits. In contrast, clin_prs_model models showed minimal performance variations across both conditions, suggesting genetic factors provide a more enduring risk background. As a static genetic burden indicator, PRS itself typically demonstrates limited predictive power. For schizophrenia models, most models maintained stable performance over extended follow-up, though 3 models outperformed at 10 years follow-up, likely due to insufficient patient numbers that limited predictive accuracy at 5 years follow-up. These results emphasize the need for future psychiatric risk prediction models to account for biomarker temporal dynamics and develop disease-specific predictive strategies.

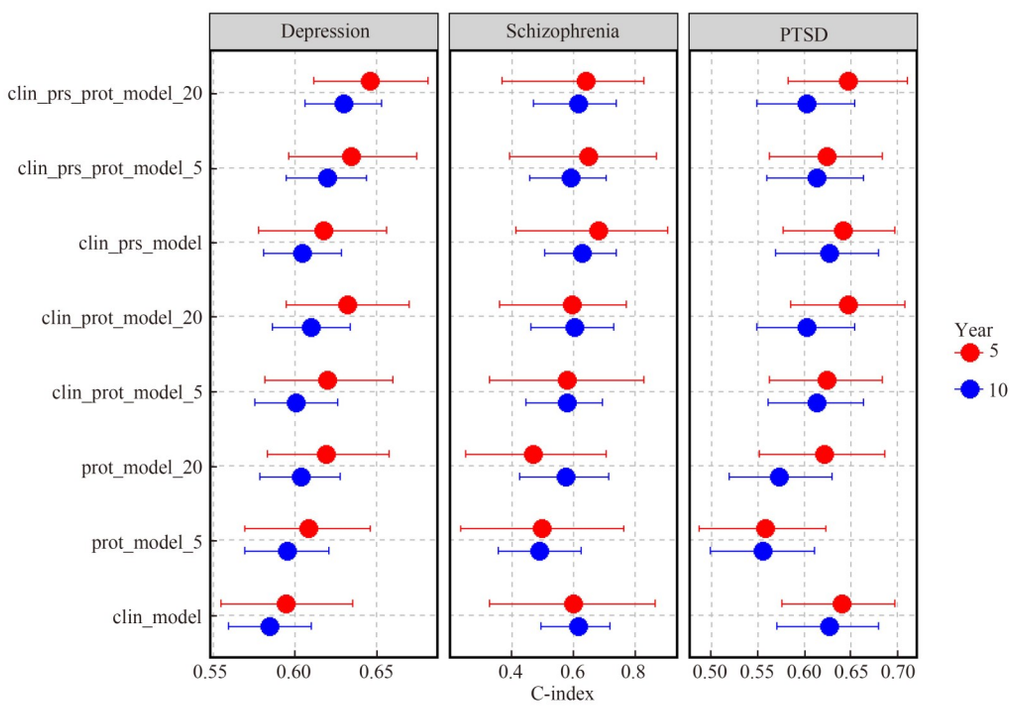

Figure 6 Results of the mental disorders risk models at 5 and 10 years of follow-up

2.6 Correlation Analyses and Enrichment Analyses

To ensure the robustness of the predictive model and gain deeper insights into its biological foundations, we conducted 2 critical analyses on the screened plasma proteins. Firstly, we evaluated potential multicollinearity issues among proteins. Correlation analyses revealed that most protein pairs exhibited low to moderate correlations ($|r|<0.5$), indicating they provided relatively independent information and effectively mitigated concerns about model instability caused by high multicollinearity. Building on this, we performed GO functional and KEGG pathway enrichment analyses. The results demonstrated that these proteins were not randomly combined but significantly enriched in several biologically significant pathways, including "positive regulation of nervous system development" and "mitochondrial transport along microtubules". These findings not only validated the biological rationality of our protein characteristics, but also provided novel molecular-level insights into the pathophysiological mechanisms of mental disorders.

3 Discussion

This study aimed to explore the predictive efficacy of plasma proteomics characteristics and PRS for 3 major mental disorders: depression, schizophrenia, and PTSD, while evaluating their clinical translation potential. Multi-omics integration has demonstrated significant value in predicting and stratifying risks of mental disorders. In de-

pression, the combination of plasma proteins and PRS significantly improved predictive performance (C-index increased by 0.038 1), achieving a prediction accuracy of 0.632 2 (95%*CI* 0.610 0-0.652 6). For schizophrenia, protein biomarkers showed moderate predictive power (C-index=0.603 2), while PRS failed to demonstrate significant improvement. In PTSD, neither the combined model of proteins nor PRS exhibited notable advantages (C-index=0.602 4). Notably, the risk stratification analyses showed that the integrated protein-PRS model was effective in distinguishing high-risk and low-risk groups (P<0.000 1), which was statistically significant for all 3 diseases. Longitudinal follow-up revealed that the predictive efficacy decreased over time, emphasizing the importance of short-term evaluation window (Figure 7).

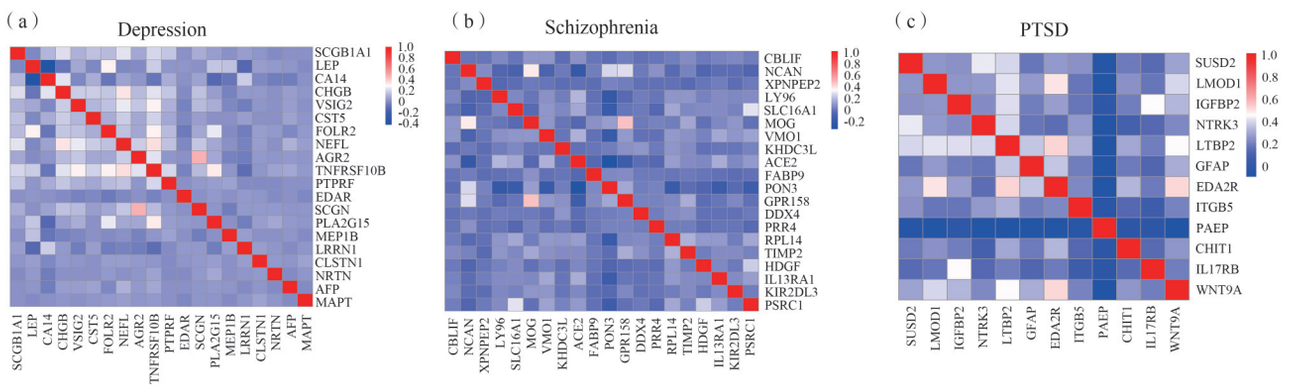

Figure 7 Correlation analysis results

Among the top 15 relatively important features in the depression risk model, 6 proteins have been confirmed by relevant studies (LEP, NEFL, SCGB1A1, SCGN, CHGB, CST5). LEP can enhance neurogenesis and neuroplasticity in the hippocampus and cortex[19], as well as modulate the hypothalamic-pituitary-adrenal axis (HPA axis) and immune system[20]. In a study by Tavast et al. to evaluate whether clinical subtypes of depression have different endocrine and metabolic characteristics, it was found that the participants with depression had significantly higher LEP levels than the control group, indicating that LEP is a risk factor for depression[21], which is consistent with our results. NEFL is a sensitive biomarker for neuronal injury[22] and has been shown to be a reliable biomarker for various neurodegenerative, inflammatory, traumatic and vascular origins of neurological diseases[23]. Previous reports have reported elevated blood NEFL levels in patients with major depression compared to physiological states[24]. Club cell secreted protein (CC16), encoded by the *SCGB1A1* gene[25], is a natural anti-inflammatory protein. In Yu et al.'s case-control study of severe depression and multiple sclerosis, it was found that depression leads to decreased serum CC16 concentrations[26]. The expression of SCGN is mainly concentrated in the neuroendocrine axis and central nervous system, and it may be used as a tool to treat chronic stress responses caused by depression in the future[27]. CHGB is a potential biomarker of the human hippocampal pathway[28] and is involved in regulating synaptic transmission[29]. It is speculated that CHGB may be involved in the pathogenesis of depression[30]. CST5 can be localized in the nucleus and modify gene transcription[31]. It is an ultra-early biomarker of traumatic brain injury[32], and the level of CST5 in the plasma of depression patients will be significantly changed[33]. These proteins are closely related to depression and have potential predictive value.

3 of the top 15 relatively important features in the schizophrenia risk model (GPR158, NCAN, and LY96) were also found in other studies. GPR158 regulates hippocampal MF-CA3 synaptic formation by interacting with glypican 4[34], and the hippocampal CA3 circuit plays a key role in memory encoding[35]. When the expression of GPR158 is affected, it leads to discrimination memory defects[36]. NCAN risk variants may affect brain structure by altering cortical folds in the occipital and prefrontal cortices, potentially establishing disease susceptibility during neurodevelopment. Research also indicates that NCAN is involved in visual processing and top-down cognitive functions. It is well known that

both of these primary cognitive processes are impaired in schizophrenia patients[37]. LY96, also known as MD2, serves as an adjuvant receptor for toll-like receptor 4 (TLR4)[38]. The TLR4-MD2 pathway may drive the progression of schizophrenia through multiple mechanisms[39], such as chronic neuroinflammation and oxidative stress.

While the integrated plasma protein model failed to significantly improve the prediction of PTSD, it is noteworthy that IGFBP2, one of our key research features, has been shown in the literature to potentially reverse the effects of acute and chronic stress and is associated with PTSD. This indicates that this protein holds significant value for future research[40]. Unfortunately, we only identified 1 validated protein. This suggests that directly selecting features from 2 920 proteins may not accurately identify those significantly associated with mental disorders. Other key proteins in the model still require further investigation to determine their potential impact. This may explain why proteins didn't show significant effects in our schizophrenia and PTSD risk models.

The genetic structure of mental disorders is highly polygenic, with thousands of common genetic variants affecting disease risk, each having a small effect. PRS is considered to be the most promising tool for summarizing the cumulative burden of genetic risk variants by summarizing individual risk alleles and returning a single estimate of an individual's genetic risk[41]. However, PRS still has significant limitations. While the weights for calculating PRS are typically determined through GWAS, GWAS research targeting mental disorders remains significantly underdeveloped, resulting in PRS calculations that fail to meet clinical practical needs[42]. Genetic factors contribute only part of the risk, and PRS can only capture part of the genetic contribution. Therefore, at this stage, PRS cannot establish or clearly predict the diagnosis of common complex diseases (e.g., mental health disorders)[10]. This may be why PRS only showed significant predictive power for depression in our study, but not for schizophrenia and PTSD.

This study demonstrates several advantages. Firstly, as a large-scale proteomics initiative, it systematically integrated plasma proteins and PRS to develop predictive models for assessing risk of depression, schizophrenia, and PTSD, providing new insights into the biological mechanisms of mental disorders. Secondly, this study conducted feature screening in large-scale protein data, and in the process of model construction, not only LASSO-Cox regression was adopted, but also 6 machine learning methods were introduced for comprehensive analysis, so as to enhance the robustness and credibility of the results. Thirdly, individualized risk prediction is the future of early detection. The prediction model constructed in this study can accurately stratify the risk of the population, and help to focus limited resources on high-risk groups. However, there are several key limitations to our study. Firstly, there is a bias in data selection because both plasma protein and cohort data are from UKB and have not been externally validated. Secondly, we subjectively selected fixed clinical variables for different mental disorders, with some showing limited effectiveness. Thirdly, to avoid population overlap, we used GWAS from Finnish databases for calculating PRS in schizophrenia and PTSD cases, which may result in underestimating genetic contributions.

In conclusion, this study provides an accurate scientific basis for the field of psychiatry. By integrating blood proteomics with PRS analyses, we demonstrate that combining genetic risk factors with biomarkers can significantly improve early screening for specific mental disorders while enabling precise population stratification. However, the predictive capacity of psychiatric PRS remains limited, requiring larger cohort studies with greater ethnic diversity and phenotypic variations in GWAS, along with technological advancements to improve its accuracy.

Acknowledgments

We thank UK Biobank for providing the proteome dataset and psychiatric population data, the FinnGen platform for providing the GWAS dataset, and all the studies, data, software, and associated participants cited. This study accessed UKB data with application number 540454.

Code availability

Analyses were performed using R software (v4.5.1), code is available at https://github.com/DAIOKD/Predicting-future-mental-disorders-based-on-plasma-protein-and-polygenic-risk-scores.

References:

[1] Kuehn B M. WHO:Pandemic sparked a push for global mental health transformation[J]. JAMA,2022,328(1):5-7.

[2] GBD 2019 Mental Disorders Collaborators. Global,regional,and national burden of 12 mental disorders in 204 countries and territories,1990—2019:A systematic analysis for the Global Burden of Disease Study 2019[J]. Lancet Psychiatry,2022,9:137-150.

[3] Hirjak D. Multidimensional perspectives on (bio)markers:Linking clinical and biological insights across psychiatric disorders,supporting a transdiagnostic model[J]. Biomarkers in Neuropsychiatry,2025,12:100129.

[4] Comai S,Manchia M,Bosia M,et al. Moving toward precision and personalized treatment strategies in psychiatry[J]. International Journal of Neuropsychopharmacology,2025,28(5):pyaf025.

[5] Topol E J. The revolution in high-throughput proteomics and AI[J]. Science,2024,385(6716):ads5749.

[6] Suhre K,McCarthy M I,Schwenk J M. Genetics meets proteomics:Perspectives for large population-based studies[J]. Nature Reviews Genetics,2021,22(1):19-37.

[7] Bhattacharyya U,John J,Lam M,et al. Circulating blood-based proteins in psychopathology and cognition:A Mendelian randomization study[J]. JAMA Psychiatry,2025,82(5):481-491.

[8] Carrasco-Zanini J,Pietzner M,Davitte J,et al. Proteomic signatures improve risk prediction for common and rare diseases[J]. Nature Medicine,2024,30(9):2489-2498.

[9] Beydoun M A,Beydoun H A,Li Z G,et al. Alzheimer's disease polygenic risk,the plasma proteome,and dementia incidence among UK older adults[J]. GeroScience,2025,47(2):2507-2523.

[10] Murray G K,Lin T,Austin J,et al. Could polygenic risk scores be useful in psychiatry? A review[J]. JAMA Psychiatry,2021,78(2):210-219.

[11] Hafeman D M,Uher R,Merranko J,et al. Person-level contributions of bipolar polygenic risk score to the prediction of new-onset bipolar disorder in at-risk offspring[J]. Journal of Affective Disorders,2025,368:359-365.

[12] Mei C,van der Gaag M,Nelson B,et al. Preventive interventions for individuals at ultra high risk for psychosis:An updated and extended meta-analysis[J]. Clinical Psychology Review,2021,86:102005.

[13] Sun B B,Chiou J,Traylor M,et al. Plasma proteomic associations with genetics and health in the UK Biobank[J]. Nature,2023,622(7982):329-338.

[14] Mullins N,Forstner A J,O'Connell K S,et al. Genome-wide association study of more than 40 000 bipolar disorder cases provides new insights into the underlying biology[J]. Nature Genetics,2021,53(6):817-829.

[15] Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. Trans-ancestry genome-wide study of depression identifies 697 associations implicating cell types and pharmacotherapies[J]. Cell,2025,188(3):640-652.

[16] Radomyslsky Z,Kivity S,Cohen R,et al. ADHD and Parkinson's disease:Unraveling the link and implications for early intervention[J]. Journal of Affective Disorders,2025,386:119462.

[17] Zhang S Y,Sun L Y,Cai D J,et al. Development and validation of PET/CT-based nomogram for preoperative prediction of lymph node status in esophageal squamous cell carcinoma[J]. Annals of Surgical Oncology,2023,30(12):7452-7460.

[18] Beyene K M,El Ghouch A. Time-dependent ROC curve estimation for interval-censored data[J]. Biometrical Journal,2022,64(6):1056-1074.

[19] Fu X Y,Wang Y C,Zhao F Y,et al. Shared biological mechanisms of depression and obesity:Focus on adipokines and lipokines[J]. Aging,2023,15(12):5917-5950.

[20] Browning B D,Schwandt M L,Farokhnia M,et al. Leptin gene and leptin receptor gene polymorphisms in alcohol use disorder:Findings related to psychopathology[J]. Frontiers in Psychiatry,2021,12:723059.

[21] Tavast I M,Solismaa A,Lyytikäinen L P,et al. Leptin and leptin receptor gene polymorphisms and depression treatment response[J]. Acta Neuropsychiatrica,2024,37:43.

[22] Meyer M A S,Beske R P,Mølstrøm S,et al. Neurofilament light chain for prognostication after cardiac arrest-first steps towards validation[J]. Critical Care,2025,29(1):348.

[23] Hviid C V B,Benros M E,Krogh J,et al. Serum glial fibrillary acidic protein and neurofilament light chain in treatment-naïve patients with unipolar depression[J]. Journal of Affective Disorders,2023,338:341-348.

[24] Bavato F,Barro C,Schnider L K,et al. Introducing neurofilament light chain measure in psychiatry:Current evidence,opportu-

nities, and pitfalls[J]. Molecular Psychiatry, 2024, 29(8): 2543-2559.
[25] Almuntashiri S, Zhu Y, Han Y H, et al. Club cell secreted protein CC16: Potential applications in prognosis and therapy for pulmonary diseases[J]. Journal of Clinical Medicne, 2020, 9(12): 4039.
[26] Yu Y, Liang H F, Chen J, et al. Postpartum depression: Current status and possible identification using biomarkers[J]. Frontiers in Psychiatry, 2021, 12: 620371.
[27] Ma Z Y, Wan Q Q, Qin W P, et al. Effect of regional crosstalk between sympathetic nerves and sensory nerves on temporomandibular joint osteoarthritic pain[J]. International Journal of Oral Science, 2025, 17(1): 3.
[28] Tukacs V, Mittli D, Hunyadi-Gulyás É, et al. Comparative analysis of hippocampal extracellular space uncovers widely altered peptidome upon epileptic seizure in urethane-anaesthetized rats[J]. Fluids and Barriers of the CNS, 2024, 21(1): 6.
[29] Podvin S, Jones J, Kang A, et al. Human iN neuronal model of schizophrenia displays dysregulation of chromogranin B and related neuropeptide transmitter signatures[J]. Molecular Psychiatry, 2024, 29(5): 1440-1449.
[30] Song J, Ma Z L, Zhang H S, et al. Identification of novel biomarkers linking depressive disorder and Alzheimer's disease based on an integrative bioinformatics analysis[J]. BMC Genomic Data, 2023, 24(1): 22.
[31] Ferrer-Mayorga G, Alvarez-Díaz S, Valle N, et al. Cystatin D locates in the nucleus at sites of active transcription and modulates gene and protein expression[J]. The Journal of Biological Chemistry, 2015, 290(44): 26533-26548.
[32] Liu J, Liu D, Sun Q, et al. Plasma proteomic signature of neonates in the context of placental histological chorioamnionitis[J]. BMJ Paediatrics Open, 2024, 8(1): e002708.
[33] Ma S M, Li R L, Gong Q, et al. Using data-driven algorithms with large-scale plasma proteomic data to discover novel biomarkers for diagnosing depression[J]. Journal of Proteome Research, 2024, 23(9): 4043-4054.
[34] Kamimura K, Maeda N. Glypicans and heparan sulfate in synaptic development, neural plasticity, and neurological disorders[J]. Frontiers in Neural Circuits, 2021, 15: 595596.
[35] Voorn R A, Vogl C. Molecular assembly and structural plasticity of sensory ribbon synapses—A presynaptic perspective[J]. International Journal of Molecular Sciences, 2020, 21(22): 8758.
[36] Li Y D, Briguglio J J, Romani S, et al. Mechanisms of memory-supporting neuronal dynamics in hippocampal area CA3[J]. Cell, 2024, 187(24): 6804-6819.
[37] Treccani M, Maggioni L, Di Giovanni C, et al. A genome-wide association study of first-episode psychosis: A genetic exploration in an Italian cohort[J]. Genes, 2025, 16(4): 16040439.
[38] Liu J, Kang R, Tang D L. Lipopolysaccharide delivery systems in innate immunity[J]. Trends in Immunology, 2024, 45(4): 274-287.
[39] Xie W Q, Luo Z H, Xiao J, et al. Identification of biomarkers related to propionate metabolism in schizophrenia[J]. Frontiers in Psychiatry, 2025, 16: 1504699.
[40] Savvidis C, Kouroglou E, Kallistrou E, et al. IGFBP-2 in critical illness: A prognostic marker in the growth hormone/insulin-like growth factor axis[J]. Pathophysiology, 2024, 31(4): 621-630.
[41] Smeland O B, Andreassen O A. Polygenic risk scores in psychiatry—Large potential but still limited clinical utility[J]. European Neuropsychopharmacology, 2021, 51: 68-70.
[42] Smeland O B, Frei O, Dale A M, et al. The polygenic architecture of schizophrenia—Rethinking pathogenesis and nosology[J]. Nature Reviews Neurology, 2020, 16(7): 366-379.

责任编辑：刘 敏