

机器学习算法分析阿尔茨海默病血浆标志物新组合

李芳玉,王彦,李颖,贾建平
首都医科大学宣武医院,北京 100053

摘要:**目的** 基于机器学习算法筛选可预测轻度认知障碍(MCI)和阿尔茨海默病(AD)的血浆蛋白标志物组合,为AD早期诊断提供依据。**方法** 从阿尔茨海默病神经影像学计划(ADNI)数据库筛选566例受试者,采集基线血浆蛋白组学数据,采用随机森林、Lasso回归、支持向量机筛选出能有效区分MCI和AD的血浆蛋白,开展脑脊液标志物、认知量表相关性分析及孟德尔随机化分析,并在GEO、UK Biobank数据库验证,构建预测模型。**结果** 三种算法共同筛选出10种核心血浆标志物,包括载脂蛋白A-II、载脂蛋白E、转甲状腺素蛋白(TTR)等,与脑脊液标志物及认知评分呈强相关;孟德尔随机化显示TTR与AD风险呈负相关。十种标志物组合预测MCI、AD的曲线下面积(AUC)分别为0.929、0.932;GEO数据集验证AUC达0.871;UK Biobank联合年龄、性别、APOE ε4,预测平均14年AD发病的AUC为0.856。**结论** 新发现的血浆蛋白组合可有效预测MCI和AD,为AD早期筛查与预防提供无创工具。
关键词:阿尔茨海默病;机器学习;随机森林;Lasso回归;支持向量机
中图分类号:R749.1 **文献标志码:**A **文章编号:**1003-8507(2026)14-2518-06
DOI:10.20043/j.cnki.MPM.202511146

Novel combinations of plasma biomarkers for Alzheimer’s disease based on machine learning algorithms

LI Fang-yu, WANG Yan, LI Ying, JIA Jian-ping
Xuanwu Hospital, Capital Medical University, Beijing 100053, China

Abstract:**Objective** To identify a novel plasma protein biomarker panel for predicting mild cognitive impairment (MCI) and Alzheimer’s disease (AD) using machine learning algorithms, and to provide evidence for early diagnosis of AD. **Methods** A total of 566 eligible participants were enrolled from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database, and baseline plasma proteomic data were collected. Random forest, Lasso regression, and support vector machine were applied to screen plasma proteins that effectively discriminate MCI and AD. Correlation analyses with cerebrospinal fluid biomarkers and cognitive scales, and Mendelian randomization analysis were conducted. The model was validated in the GEO and UK Biobank databases. **Results** Ten core plasma biomarkers, including apolipoprotein AII, apolipoprotein E, transthyretin (TTR), were consistently identified by the three algorithms and showed strong correlations with cerebrospinal fluid biomarkers and cognitive scores. Mendelian randomization indicated an inverse association between TTR and AD risk. The area under the curve (AUC) of the tenbiomarker panel was 0.929 for predicting MCI and 0.932 for AD. The panel achieved an AUC of 0.871 in the GEO dataset. When combined with age, sex and APOE ε4 status, the panel yielded an AUC of 0.856 for predicting incident AD over an average of 14 years in the UK Biobank. **Conclusion** The newly identified plasma protein panel can effectively predict the occurrence of MCI and AD, providing a noninvasive tool for early screening and prevention of AD.
Keywords:Alzheimer’s disease; Machine learning; Random forest; Lasso regression; Support vector machine

阿尔茨海默病(Alzheimer’s disease, AD)是老年

期痴呆最主要的病因,中国60岁及以上人群AD患病率为3.9%(约983万),轻度认知障碍(mild cognitive impairment, MCI)患病率15.5%(约3877万)^[1]。由于疾病晚期难以治疗,早期诊断至关重要。目前AD经典生物标志物多依赖脑脊液检测,但其创伤较大、依从性低,难以用于大规模人群筛查。血浆标志物具有微创、易获取、可重复检测等优势,是理想的替代方案。随着蛋白质组学与公共数据库的发展,大量潜在AD血液标志物被报道^[2-3],但多数

基金项目: 国家科技部科技创新2030-重大项目子课题(2021ZD0201802);北京市科学技术委员会资助课题(Z251100004625066);昌平实验室(2025B-07-07);宣武医院院级课题(LCYJ202310)
作者简介: 李芳玉和王彦为共同第一作者。李芳玉(1988—),女,硕士,主管技师,研究方向:痴呆流行病学和标志物研究;王彦(1986—),女,博士,助理研究员,研究方向:痴呆标志物研究
通信作者: 贾建平, E-mail: jjp@ccmu.edu.cn

缺乏独立验证,未形成稳定、可推广的预测模型。机器学习可高效处理高维组学数据,在生物标志物筛选与疾病预测中优势显著^[4-6]。随机森林、Lasso 回归、支持向量机是 AD 研究中常用的三种算法。

本研究基于阿尔茨海默病神经影像学计划 (ADNI) 数据库血浆蛋白组学数据,联合三种机器学习方法筛选可同时区分与预测 MCI 和 AD 的血浆蛋白组合,并在基因表达综合数据库 (GEO) 横断面数据集、英国生物银行 (UK Biobank) 纵向随访队列中验证,构建高效、无创的早期预测模型,为 AD 早期筛查提供新方法。

1 材料与方法

1.1 数据来源 ADNI 数据库 (adni.loni.usc.edu) 是一项纵向、多中心研究。本研究筛选 566 例符合入组条件的受试者,包括 58 例认知正常 (normal cognition, NC)、396 例 MCI 和 112 例 AD,收集基线期血浆蛋白组学数据。本研究为二次数据分析,获机构伦理审查豁免。

1.2 模型的变量筛选及参数设置 以 146 种血浆蛋白为候选变量,采用三种算法进行筛选:(1) 随机森林:使用 R 包“randomForest”,决策树数量设为 1 000,以平均准确率降低 (Mean Decrease Accuracy, MDA)、平均基尼不纯度降低 (Mean Decrease Gini, MDG) 评估变量重要性,采用分层抽样平衡样本,并设置 class_weight = “balanced” 参数,根据类别样本量自动调整权重。(2) Lasso 回归:使用 R 包“glmnet”,加权 Lasso 回归平衡类别 (以样本量倒数为权重),alpha 设置为 1,二分类模型,以 lambda.min 确定最优变量数。(3) 支持向量机:使用 R 包“Caret”,径向基核函数,十折分层交叉验证,调整权重校正小样本偏倚。变量纳入标准:Lasso 回归或支持向量机筛选出的变量与随机森林两指标中筛选出的变量任一重合即可纳入模型。

1.3 模型构建与验证 核心蛋白经 Z score 标准化,采用加权 logistic 回归构建组合得分 (comprehensive score, CS),以样本量倒数为权重,避免小样本组被稀释。本研究 NC 组样本量仅 58 例,为避免划分训练/测试集后验证集统计学效力不足,本研究基于全样本完成蛋白筛选与模型构建。

分析蛋白及组合得分与脑脊液 A β 、Tau、pTau 及多项认知量表的相关性;采用逆概率处理加权进行孟德尔随机化,分析蛋白与 AD 的因果关联;在 GEO (GSE122063) 数据集 (包含 56 例 AD 患者与 44 名 NC 的脑组织样本) 进行横断面验证;在 UK Biobank 具有基线血浆蛋白组学数据的 53 044 例参与者 (平均随访 14.17 年) 中进行纵向发病风险预测验证 (项目

ID:150493)。

1.4 统计方法 采用 R 4.2.2 软件 (<https://www.r-project.org/>) 分析。计量资料以 $\bar{x} \pm s$ 表示,计数资料以例数 (%) 表示;分类变量组间比较采用 χ^2 检验、连续变量采用单因素方差分析;相关性采用 Pearson 或 Spearman 分析;预测效能采用受试者工作特征曲线下面积 (AUC) 评估,双侧检验水准 $\alpha = 0.05$ 。

2 结果

2.1 研究对象基本情况 本研究基线期共纳入 566 名研究对象,其中 NC 组 58 例, MCI 组 396 例, AD 组 112 例。平均年龄分别为 (75.11 $\pm$ 5.77) 岁、(74.74 $\pm$ 7.40) 岁、(74.80 $\pm$ 8.07) 岁;女性占比分别为 48.3%、35.4% 和 42.0%。

2.2 三种机器学习方法筛选 AD 预测变量结果

(1) 利用随机森林对 146 个蛋白变量进行筛选,基于 MDA 和 MDG 进行重要性排序,两指标重要性排名前 20 的变量情况见图 1。

(2) Lasso 回归分析,变量系数绝对值排名前 20 的变量分别为 IL-16、A2Macro、Apo A-II、Apo E、Apo B、IL-6r、Vitronectin、MIP-1 alpha、TTR、CLU、BNP、MIG、PYY、PPP、MMP-1、IL-3、CEA、RAGE、MIP-1 beta 和 Cortisol。

(3) 支持向量机排名前 20 的变量分别为 SGOT、Apo A-II、BNP、Apo E、Eotaxin-3、A1Micro、IGM、PYY、BTC、PLGF、PAPP-A、TTR、TN-C、A2Macro、MIP-1 alpha、PPP、Vitronectin、IL-16、TIMP-1、HB-EGF。

最终选取随机森林与 Lasso 回归或支持向量机有重合的蛋白,能共同预测 AD 的变量具体为: A1Micro、A2Macro、Apo A-II、Apo E、BTC、BNP、Eotaxin-3、IGM、IL-16、MIP-1 alpha、PPP、PYY、PLGF、PAPP-A、TN-C、TTR、SGOT、RAGE。

2.3 三种机器学习方法筛选 MCI 预测变量结果

(1) 随机森林对 146 个蛋白变量进行筛选,并按 MDA 和 MDG 进行重要性排序,两指标重要性排名前 20 的变量情况见图 2。

(2) Lasso 回归分析,变量系数绝对值排名前 20 的变量分别为 MIP-1 alpha、TTR、Serotransferrin、Apo D、PARC、Apo A-II、Apo E、Osteopontin、Calcitonin、IL-16、Apo A-IV、MCP-4、HB-EGF、PAPP-A、ACE、Nr-CAM、ANG-2、Apo B、FasL 和 AACT。

(3) 支持向量机排名前 20 的变量分别为 Apo E、Apo A-II、TTR、MIP-1 alpha、HB-EGF、BNP、FasL、Eotaxin-3、IL-16、Apo A-IV、PYY、ANG-2、IGM、BTC、PAPP-A、Vitronectin、Leptin、Apo D、CRP、MMP

- 10。

最终选取随机森林与 Lasso 回归或支持向量机有重合的蛋白,能共同预测 MCI 的变量具体为:ANG -

2、Apo A - II、Apo D、Apo E、BTC、BNP、Calcitonin、Eotaxin - 3、FasL、HB - EGF、IL - 16、MIP - 1 alpha、PYY、PAPP - A、Serotransferrin、TTR。

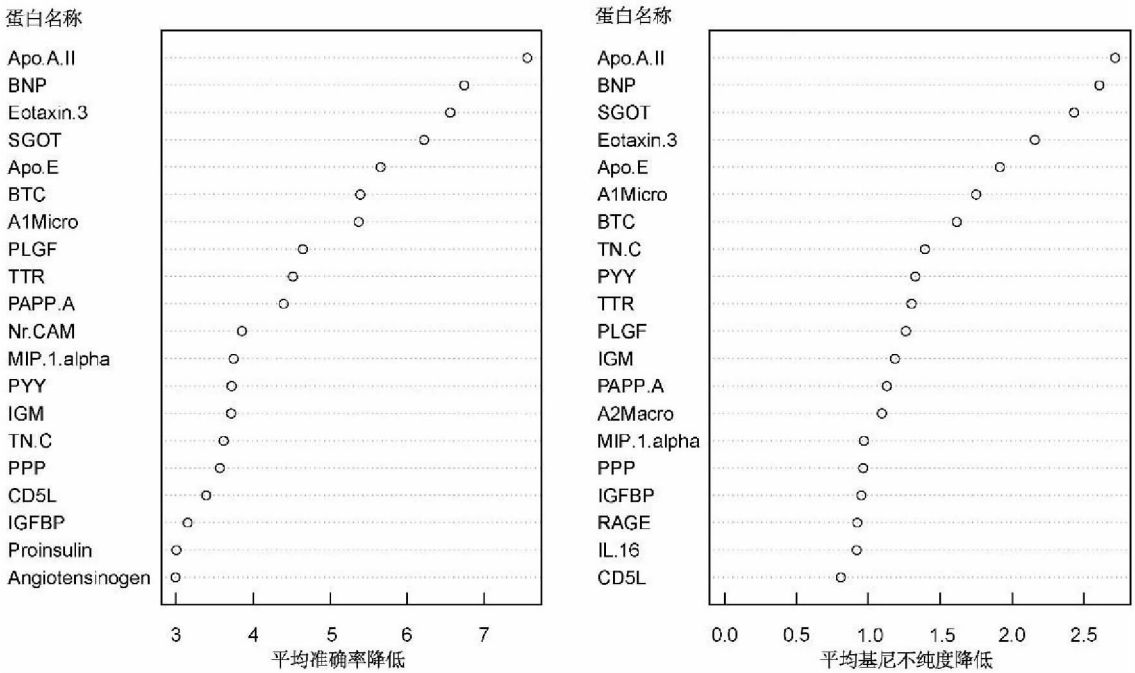


图 1 随机森林方法筛选能预测 AD 的前 20 名蛋白

Fig.1 The top 20 variables selected by random forest for AD

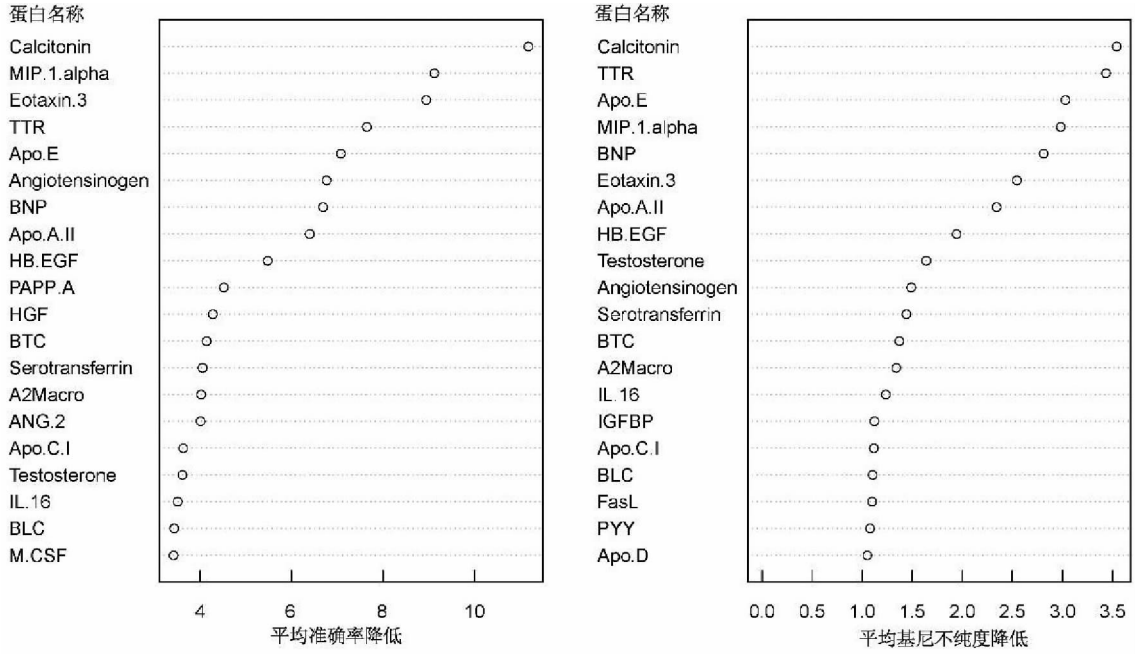


图 2 随机森林方法筛选能预测 MCI 的前 20 名蛋白

Fig.2 The top 20 variables selected by random forest for MCI

2.4 构建预测模型 在至少两种机器学习算法中被识别为重要变量的蛋白,结合研究对象基本信息,最终能同时预测 MCI 和 AD 的变量是:Apo A - II、Apo

E、BTC、BNP、Eotaxin - 3、MIP - 1 alpha、PYY、PAPP - A、TTR、IL - 16。进一步通过随机森林算法对这十个变量进行重要性排序,对于 AD 和 MCI 来说,重要性

排名前五的变量中有三个相同变量,分别是:Apo E、MIP-1 alpha、Apo A-II。

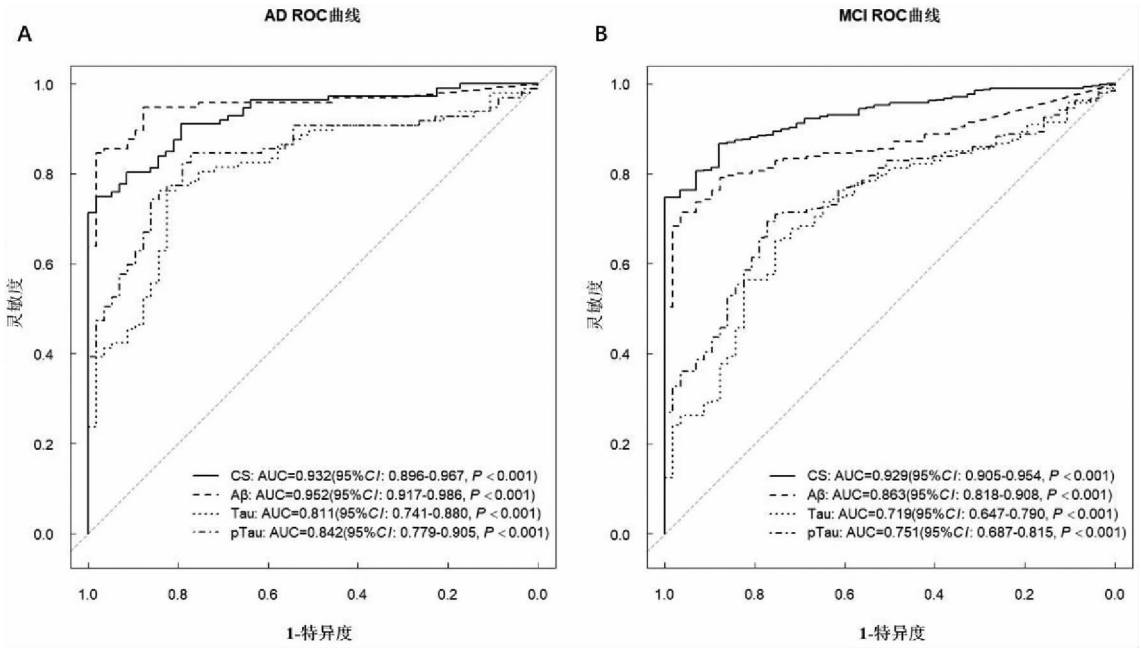
进一步将这十个变量进行标准化,通过加权 logistic 回归形成一个组合得分,对于 AD 来说:

$$CS = 1.087 - 0.779 \times Apo\ A-II - 1.191 \times Apo\ E + 0.777 \times BTC + 1.327 \times BNP + 0.018 \times Eotaxin - 3 - 1.514 \times IL-16 + 0.627 \times MIP-1\ alpha + 0.916 \times PYY - 0.596 \times PAPP-A - 0.336 \times TTR;$$

对于 MCI 来说:

$$CS = 1.917 - 0.847 \times Apo\ A-II - 1.086 \times Apo\ E + 0.417 \times BTC + 0.399 \times BNP + 0.336 \times Eotaxin - 3 - 1.350 \times IL-16 + 0.694 \times MIP-1\ alpha + 0.855 \times PYY + 0.411 \times PAPP-A - 0.721 \times TTR.$$

为了避免蛋白名称中的数字与公式数值混淆,公式中蛋白名称改用斜体。该组合预测 AD 和 MCI 的曲线下面积(95% CI)分别为:0.932(0.896~0.967)和0.929(0.905~0.954)。对于 AD 来说,这十个蛋白的组合预测得分与脑脊液 Aβ 预测能力接近,甚至超过了脑脊液 Tau 和 pTau。对于 MCI 来说,十个蛋白的综合得分预测能力比脑脊液 Aβ、Tau 和 pTau 都高(图3)。受 NC 组小样本限制,指标未经过独立测试集验证,本研究进行了 50 次分层随机抽样模拟验证,结果显示,AD 组平均 AUC 为 0.934,标准差为 0.030;MCI 组平均 AUC 为 0.923,标准差为 0.027,与全样本结果基本一致。



注:CS 表示十个蛋白的组合得分,下同。

图3 十个蛋白综合得分及脑脊液 Aβ、Tau、pTau 预测 AD(A)和 MCI(B)受试者工作特征曲线

Fig.3 Receiver operating characteristic curves of composite scores, CSF Aβ, CSF Tau, and CSF pTau for predicting AD (A) and MCI (B); Note: CS: Comprehensive score

2.5 临床相关性分析结果 将上述十个蛋白及其综合得分分别与脑脊液 Aβ、Tau、pTau 和量表得分做相关性分析,结果发现对于 AD 患者,CS、Apo E、Eotaxin-3 以及 Apo A-II 与脑脊液 Aβ 有较高的相关性($P < 0.05$)。除此之外,CS 与 CDRSB、ADAS-Q4 和 MMSE 都有很强的相关性($P < 0.05$)。在 MCI 患者中,也呈现出相似趋势(图4)。

2.6 孟德尔随机化结果 孟德尔随机化分析,十个蛋白中除 Apo E 外,仅 TTR 的因果关联得到验证($P = 0.037$)。逆方差加权提示 TTR 与 AD 风险呈负相关,Q 统计量表明分析中使用的工具变量之间无异质

性($P > 0.05$)。

2.7 模型验证结果 (1)GEO GSE122063 脑组织数据中,十个蛋白综合得分预测 AD 的曲线下面积达到 0.871(0.801~0.942),Apo E、BTC、IL-16、MIP-1 alpha、PAPP-A 相关基因表达在 AD 与 NC 组之间差异有统计学意义($P < 0.05$)。(2)UK Biobank 随访数据库 53 044 例受试者中,在随访平均 14.17 年后,有 731 例发展为 AD,该蛋白组数据于基线期采集,其中九种蛋白(除 BTC 外)基线水平在 AD 组与 NC 组间差异有统计学意义(校正后 P 值均小于 0.05)。十个蛋白组合得分联合性别、年龄、APOE 预测平均 14 年

后发生 AD 的曲线下面积达到 0.856(0.838 ~ 0.873) (图 5A)。进一步将组合得分按 1: 1 分成高低两组, 发现两组发生 AD 的风险差异有统计学意义 ($P < 0.001$), 特别是 70 岁以后(图 5B)。

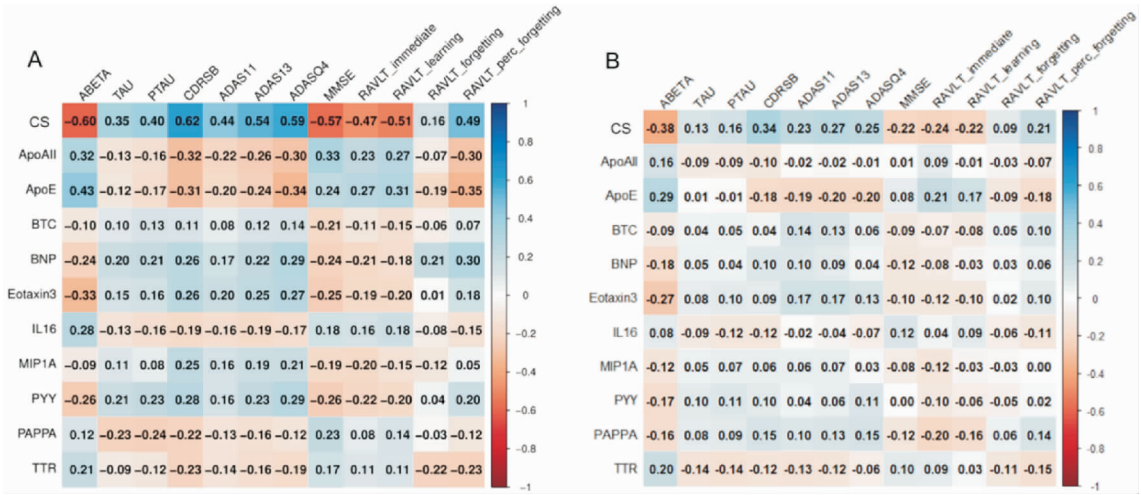
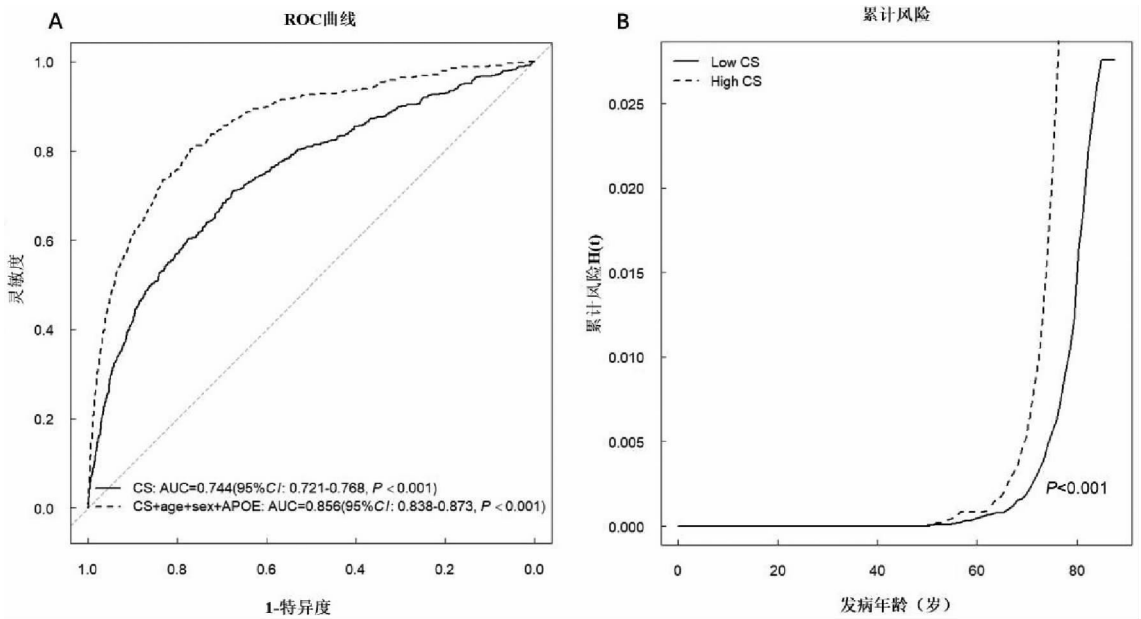


图 4 十个蛋白及综合得分与脑脊液 Aβ、Tau、pTau 以及其他临床量表得分在 AD 组(A)和 MCI 组(B)相关性结果

Fig. 4 Correlation results of the ten proteins and their composite scores with CSF Aβ, CSF tau, CSF pTau, and other clinical scale scores in the AD group (A) and the MCI group (B)



注: High CS 和 Low CS 分别表示 CS 高、低得分组。

图 5 十个蛋白综合得分在 UK Biobank 随访数据中预测 AD 受试者工作特征曲线(A)以及不同综合得分发生 AD 的累计风险(B)

Fig. 5 Receiver operating characteristic curve of the composite score for predicting AD in the UK Biobank follow-up data (A) and cumulative risk of AD (B); Note: High CS, High comprehensive score; Low CS, Low comprehensive score

3 讨论

本研究联合随机森林、Lasso 回归、支持向量机三种机器学习算法,基于 ADNI 数据库,从 146 种血浆蛋白中筛选出 10 种核心蛋白,构建可同时预测 MCI 与

AD 的血液标志物模型,经多队列验证效能稳定,为 AD 早期无创筛查提供可靠工具。

随着 AD 多组学数据库的出现,机器学习算法在 AD 分类与预测中的应用越来越广泛。王凤琳等采用随机森林、Lasso 回归、XGBoost 算法对 ADNI 数据库

中的脂蛋白及代谢物进行变量重要性排序及筛选,发现了对 AD 发病起着重要影响的脂蛋白^[6]。赵彩丽等利用 Lasso 回归、EN、随机森林、XGBoost 等机器学习算法在不同 SNPs 阈值下对 AD 遗传风险预测进行建模,探索了 AD 遗传风险预测模型统计建模新策略^[7]。本研究综合三种机器学习方法,可降低单一机器学习方法筛选变量所致结果偏倚较大的风险,进而提升后续建模的准确度。

通过随机森林对筛选出的十个变量进行重要性排序,发现 Apo E 和 Apo A - II 对 MCI 和 AD 来说都比较重要。Apo E 参与 β - 淀粉样蛋白清除和脂质转运,此前也有研究报道 AD 组血浆 Apo E 水平显著下降^[8]。血浆和中枢神经系统中 Apo E 水平降低与 AD 的发生有关^[9]。Apo E 还参与胆碱能通路,维持脂质稳态,这对于乙酰胆碱的合成至关重要^[10-11]。多项临床研究发现,AD 患者血浆中 Apo A - II 水平显著低于非 AD 对照组,且其下降幅度甚至大于 Apo A - I。这种变化可能与 AD 的病理过程相关,提示 Apo A - II 可能作为 AD 的潜在血液生物标志物^[12]。孟德尔随机化还发现 TTR 与 AD 呈负相关,可能是 AD 的保护因素。此前有研究报道 TTR 在 AD 患者脑脊液中水平降低^[13],也有研究发现 TTR 是血浆中可以区分快速进展与缓慢进展 AD 患者的关键蛋白之一^[14]。

本研究局限性:NC 组样本量较小,未严格分离训练/测试集,虽经 50 次随机抽样模拟验证,仍存在潜在信息泄漏风险;未收集共病及用药史信息,可能导致标志物筛选偏差、降低模型外推适用性;预测模型主要基于欧美人群,尚未在中国人群中验证。后续将扩大样本量,采用严格交叉验证,并开展中国人群外部验证,构建更适合国人的预测模型。

综上所述,本研究发现的包括 Apo A - II、Apo E、TTR 在内的十种血浆蛋白标志物及其组合能有效鉴别和预测 MCI 及 AD,可为 AD 早期预警、高危人群筛查及防控干预提供新思路与理论依据。

利益冲突声明 本研究不存在任何利益冲突

参考文献

[1] Jia LF, Du YF, Chu L, et al. Prevalence, risk factors, and management of dementia and mild cognitive impairment in adults aged 60 years or older in China; a cross - sectional study [J]. *Lancet Public Health*, 2020, 5(12): e661 - e671.

[2] Jiang Y, Zhou X, Ip FC, et al. Large - scale plasma proteomic profiling identifies a high - performance biomarker panel for Alzheimer's disease screening and staging [J]. *Alzheimer's & Dementia*, 2021, 18(1): 88 - 102.

[3] Bai B, Vanderwall D, Li YX, et al. Proteomic landscape of

alzheimer's disease: novel insights into pathogenesis and biomarker discovery [J]. *Molecular Neurodegeneration*, 2021, 16(1): 55.

[4] Yu ZW, Mulholland A, Huang TY, et al. Multimodal AI for alzheimer disease diagnosis: systematic review of datasets, models, and modalities [J]. *Journal of Medical Internet Research*, 2026, 28(3): e85414.

[5] Jia M, Wu Y, Xiang C, et al. Predicting alzheimer's disease with interpretable machine learning [J]. *Dementia and Geriatric Cognitive Disorders*, 2023, 52(4): 249 - 257.

[6] 王凤琳,王爱民,黄一铭,等. 利用机器学习算法分析影响阿尔兹海默病的脂蛋白及代谢物 [J]. *现代预防医学*, 2023, 50(23): 4225 - 4230.

Wang FL, Wang AM, Huang YM, et al. Analysis of lipoproteins and metabolites affecting alzheimer's disease by machine learning algorithm [J]. *Modern Preventive Medicine*, 2023, 50(23): 4225 - 4230. (In Chinese)

[7] 赵彩丽,陈帅,王荣,等. 不同 SNPs 阈值下机器学习与多基因风险评分的阿尔茨海默症遗传风险预测研究 [J]. *现代预防医学*, 2023, 50(3): 396 - 401, 407.

Zhao CL, Chen S, Wang R, et al. Genetic risk prediction of alzheimer's disease based on machine learning and polygenic risk score under different SNPs thresholds [J]. *Modern Preventive Medicine*, 2023, 50(3): 396 - 401, 407. (In Chinese)

[8] Zhuang TC, Yang YQ, Ren HL, et al. Novel plasma protein biomarkers; a time - dependent predictive model for alzheimer's disease [J]. *Archives of Gerontology and Geriatrics*, 2025, 129(2): 105650.

[9] Wijesinghe P, Li HR, Ai ZY, et al. Apolipoprotein E dysfunction in Alzheimer's disease: a study on miRNA regulation, glial markers, and amyloid pathology [J]. *Front Aging Neurosci*, 2024, 16: 1495615.

[10] Chen J, Zhao S, Zhou Y, et al. APOE4 reprograms microglial lipid metabolism in alzheimer's disease: mechanisms and therapeutic implications [J]. *Bioscience Trends*, 2025, 19(6): 641 - 658.

[11] Soares HD. Hypercholesterolemia and alzheimer's disease: unraveling the connection and assessing the efficacy of lipid - lowering therapies [J]. *Journal of Alzheimer's Disease*, 2024, 101(s1): S371 - S393.

[12] 张芳,王晓良. 血液载脂蛋白作为神经退行性疾病潜在生物标志物的研究 [J]. *药学报*, 2021, 56(6): 1513 - 1520.

Zhang F, Wang XL. Study on blood apolipoprotein as a potential biomarker of neurodegenerative diseases [J]. *Acta pharmaceutica Sinica*, 2021, 56(6): 1513 - 1520. (In Chinese)

[13] Corino C, Aimo A, Luigetti M, et al. Tetrameric transthyretin as a protective factor against alzheimer's disease [J]. *Molecular Neurobiology*, 2025, 62(3): 2945 - 2954.

[14] Tien YT, Lee WJ, Liao YC, et al. Plasma transthyretin as a predictor of amnesic mild cognitive impairment conversion to dementia [J]. *Scientific Reports*, 2019, 9(1): 18691.

收稿日期:2025-11-11