

基于机器学习筛选及孟德尔随机化的实验室指标与 1 型糖尿病因果关联研究^{*}

李 贇¹, 杜 佳¹, 田肃岩², 关贺元¹, 孙 琪¹, 许建成^{1△}

吉林大学第一医院: 1. 检验科; 2. 临床研究部, 吉林长春 130021

摘要: **目的** 基于机器学习筛选潜在关联的实验室指标, 并联合孟德尔随机化分析探讨其与 1 型糖尿病 (T1DM) 的因果关联。 **方法** 选择 2021—2022 年于吉林大学第一医院就诊的 T1DM 患者 210 例作为 T1DM 组, 选择与 T1DM 组年龄、性别匹配的同期在吉林大学第一医院体检的健康体检者 210 例作为健康对照组。按 7: 3 的比例将全部研究对象划分为训练集 ($n=294$) 与测试集 ($n=126$)。采用 Lasso 回归结合交叉验证从 39 项实验室指标中筛选指标, 构建极端梯度提升 (XGBoost)、Logistic 回归、轻量梯度提升机 (LightGBM)、随机森林 (RF)、自适应提升 (AdaBoost)、决策树 (DT)、梯度提升决策树 (GBDT)、支持向量机 (SVM) 8 种机器学习模型, 绘制受试者工作特征 (ROC) 曲线比较 8 种机器学习模型的性能; 基于英国生物银行及日本生物银行的单核苷酸多态性 (SNP) 数据, 分别在欧洲人群和亚洲人群中分析所选指标与 T1DM 的因果关联。 **结果** Lasso 回归在 39 项实验室指标中筛选出 10 项候选特征: 肌酸激酶 (CK)、天冬氨酸氨基转移酶 (AST)、总蛋白 (TP)、白蛋白 (ALB)、间接胆红素 (IBil)、视黄醇结合蛋白 (RBP)、铁蛋白 (SF)、总铁结合能力 (TIBC)、不饱和铁结合能力 (UIBC) 和血红蛋白 (Hb)。基于 10 项候选特征, 构建 8 种机器学习模型, 其中 XGBoost 模型在测试集中的 AUC 最大 (0.968, 95%CI: 0.945~0.991), 且训练集与测试集 AUC 差值最小 ($\Delta AUC=0.020$)。基于 XGBoost 模型的特征重要性分析, 10 项指标按重要性从高到低依次为 RBP、TIBC、TP、ALB、UIBC、CK、Hb、SF、IBil 和 AST。孟德尔随机化分析结果显示, ALB 与欧洲人群 T1DM 之间存在潜在因果关联 ($P<0.05$), 使用随机效应逆方差加权法 ($OR=0.900$, 95%CI: 0.835~0.970, $P=0.006$)、加权中位数法 ($OR=0.896$, 95%CI: 0.836~0.961, $P=0.002$) 和加权模式法 ($OR=0.900$, 95%CI: 0.832~0.974, $P=0.010$) 进行分析得到了 186 个能够显著影响 ALB 与 T1DM 因果关系的 SNP。MR-Steiger 方向性检验结果支持因果方向为 ALB→T1DM。在亚洲人群中, 受限于 GWAS 样本量及工具变量数量, ALB 与 T1DM 的因果关联无统计学意义 ($P>0.05$)。 **结论** 该研究成功构建并验证了整合机器学习筛选与孟德尔随机化验证的分析框架。该策略能有效从常规临床指标中识别出与 T1DM 具有因果关联的风险因子 ALB, 并排除非因果性的伴随指标, 为临床大数据的因果挖掘提供了新思路。

关键词: 1 型糖尿病; 白蛋白; 机器学习; 孟德尔随机化; 因果关联

中图法分类号: R587.1; R195.1; R446

文献标志码: A

文章编号: 1672-9455(2026)16-2161-12

Causal associations between laboratory parameters and type 1 diabetes based on machine learning combined with Mendelian randomization^{*}

Li Ze¹, Du Jia¹, Tian Suyan², Guan Heyuan¹, Sun Qi¹, Xu Jiancheng^{1△}

1. Department of Laboratory Medicine; 2. Division of Clinical Research, the First Hospital of Jilin University, Changchun, Jilin 130021, China

Abstract: Objective To identify potential laboratory biomarkers associated with type 1 diabetes (T1DM) based on machine learning, and to explore their causal associations with T1DM through Mendelian randomization analysis. **Methods** A total of 210 T1DM patients, who were treated at the First Hospital of Jilin University from 2021 to 2022, were selected as the T1DM group, and 210 healthy subjects, who underwent physical

^{*} 基金项目: 吉林省医疗卫生人才专项 (JLSRCZX2025-036)。

作者简介: 李贇, 男, 硕士研究生在读, 主要从事基于检验大数据使用机器学习算法建立疾病预测模型研究。△ 通信作者, E-mail: xjc@jlu.edu.cn。

网络首发 [https://link.cnki.net/urlid/50.1167.r.20260727.1826.008\(2026-07-28\)](https://link.cnki.net/urlid/50.1167.r.20260727.1826.008(2026-07-28))

中文引用格式: 李贇, 杜佳, 田肃岩, 等. 基于机器学习筛选及孟德尔随机化的实验室指标与 1 型糖尿病因果关联研究[J]. 检验医学与临床, 2026, 23(16): 2161-2172.

英文引用格式: Li Z, Du J, Tian S Y, et al. Causal associations between laboratory parameters and type 1 diabetes based on machine learning combined with Mendelian randomization[J]. Lab Med Clin, 2026, 23(16): 2161-2172.

examinations at the First Hospital of Jilin University during the same period and were matched with the T1DM group in terms of age and gender, were selected as the healthy control group. All the subjects were divided into a training set ($n=294$) and a test set ($n=126$) at a ratio of 7 : 3. Lasso regression combined with cross-validation was used to screen candidate features from 39 laboratory indicators. Eight machine-learning models, including extreme gradient boosting (XGBoost), Logistic regression, light gradient boosting machine (LightGBM), random forest (RF), adaptive boosting (AdaBoost), decision tree (DT), gradient boosting decision tree (GBDT) and support vector machine (SVM), were constructed, and receiver operating characteristic (ROC) curves were drawn to compare the performance of the eight machine-learning models. Based on the single-nucleotide polymorphism (SNP) data from the UK Biobank and the Japanese Biobank, the causal associations between the selected indicators and T1DM were analyzed in European and Asian populations, respectively. **Results** Lasso regression screened out 10 candidate features from the 39 laboratory indicators: creatine kinase (CK), aspartate aminotransferase (AST), total protein (TP), albumin (ALB), indirect bilirubin (IBil), retinol-binding protein (RBP), ferritin (SF), total iron-binding capacity (TIBC), unsaturated iron-binding capacity (UIBC) and hemoglobin (Hb). Based on the 10 candidate features, eight machine learning models were developed. The XGBoost model achieved the highest area under the ROC curve (AUC) in the test set (0.968, 95%CI: 0.945–0.991), and exhibited the smallest difference in AUC between the training set and the test set ($\Delta\text{AUC}=0.020$). Feature importance analysis based on the XGBoost model showed that the 10 indicators in descending order of importance were as follows: RBP, TIBC, TP, ALB, UIBC, CK, Hb, SF, IBil and AST. The results of Mendelian randomization analysis revealed a potential causal association between ALB and T1DM in European populations ($P<0.05$). Using the random-effects inverse-variance weighted method ($OR=0.900$, 95%CI: 0.835–0.970, $P=0.006$), weighted median method ($OR=0.896$, 95%CI: 0.836–0.961, $P=0.002$) and weighted mode method ($OR=0.900$, 95%CI: 0.832–0.974, $P=0.010$), 186 SNP significantly influencing the causal relationship between ALB and T1DM were identified. The results of the MR-Steiger directional test supported the causal direction from ALB to T1DM. In Asian populations, due to limited sample size in GWAS and the number of instrumental variables, no statistically significant causal association between ALB and T1DM was observed ($P>0.05$). **Conclusion** This study successfully constructs and validates an analytical framework integrating machine-learning-based screening and Mendelian randomization. This strategy can effectively identify causally related risk factor ALB associated with T1DM from routine clinical indicators and exclude non-causal confounders, offering a novel strategy for causal discovery in clinical big data.

Key words: type 1 diabetes; albumin; machine learning; Mendelian randomization; causal association

1 型糖尿病 (T1DM) 是一种代谢紊乱综合征, 其特征是由于胰岛素绝对缺乏引起的高血糖。T1DM 主要是由免疫介导的胰岛 β 细胞破坏所致。近几十年来, 全球范围内 T1DM 发病率持续增加, 并且显著存在地区差异性。如欧洲国家芬兰、德国和挪威的发病率每年分别增加 2.4%、2.6% 和 3.3%^[1-3], 而亚洲国家如中国、印度等并不显著^[4]。

T1DM 患者的部分临床实验室指标与健康人群存在显著差异。以往研究已报道 T1DM 患者存在多项实验室指标异常, 如肝酶紊乱伴血糖及血脂代谢异常^[5]、肾小球滤过率下降^[6]、糖化血红蛋白水平升高^[7]以及血清镁水平降低^[8]等。机器学习是一种高度智能的工具, 能够从海量数据中筛选出与目标疾病显著相关的指标。然而, 目前尚缺乏研究基于肝功能、肾功能、微量元素、血红蛋白、血小板等多类临床实验室指标, 采用机器学习方法筛选与 T1DM 显著相关的指标。

通过免疫调节或基因治疗预防 T1DM 的尝试尚未取得成功。孟德尔随机化 (MR) 基于孟德尔遗传定律, 利用与暴露强相关的遗传变异作为工具变量, 是推断疾病与相关暴露因素间因果关联的有力工具。与传统技术相比, MR 具有正交假设特性, 且由于疾病状态通常不会改变生殖细胞的 DNA 序列, 反向因果关系的发生概率较低^[9]。更为重要的是, 由于减数分裂过程中基因型呈随机分配, MR 可有效控制混杂因素^[10]。已有 MR 分析表明, 维生素 D 水平低下可能增加 T1DM 易感性^[11]; 此外, 嗜酸性粒细胞水平升高是 T1DM 发生的危险因素^[12]; 与丙氨酸氨基转移酶 (ALT) 相关的非酒精性脂肪性肝病 (NAFLD) 对早期 T1DM 肾病及 T1DM 酮症酸中毒具有因果效应^[13]。由此可见, 众多临床实验室指标与 T1DM 之间可能存在因果关联, 需进一步筛选与深入探讨。机器学习是挖掘疾病相关差异指标的强大工具, 而 MR 分析则是推断疾病与相关因素间因果关联的精准手

段。本研究提出将机器学习与 MR 相结合的技术方案:首先利用机器学习从海量指标中筛选目标因素,随后通过 MR 从因果角度验证这些因素与亚洲和欧洲人群中 T1DM 发生风险的关联,旨在为 T1DM 的防治提供参考。此外,该方法可为临床寻找疾病相关风险因素提供新思路,通过快速精准定位风险因素,提升研究效率并惠及临床患者。现报道如下。

1 资料与方法

1.1 一般资料 选择 2021—2022 年于吉林大学第一医院就诊的 T1DM 患者 210 例作为 T1DM 组。纳入标准:符合美国糖尿病协会的现行标准并被确诊为 T1DM^[14]。排除标准:(1)合并恶性肿瘤、血液系统疾病;(2)合并感染性疾病;(3)患有除 T1DM 外的其他自身免疫性疾病或严重代谢性疾病。同时选择与 T1DM 组年龄、性别匹配的同期在吉林大学第一医院体检的健康体检者 210 例作为健康对照组,要求排除患恶性肿瘤、血液系统疾病、感染性疾病、代谢性疾病、自身免疫性疾病,且常规体格检查及影像学检查结果均未见明显异常。T1DM 组中男 86 例,女 124 例;男性中位年龄 9(7,12)岁、体质指数(BMI)为 $(21.8 \pm 2.7) \text{ kg/m}^2$;女性中位年龄 11(8,14)岁、BMI 为 $(19.2 \pm 3.1) \text{ kg/m}^2$ 。健康对照组中男 86 例,女 124 例;男性中位年龄 10(7,14)岁、BMI 为 $(21.5 \pm 3.1) \text{ kg/m}^2$;女性中位年龄 11(8,14)岁、BMI 为 $(19.2 \pm 2.7) \text{ kg/m}^2$ 。2 组性别、年龄、BMI 比较,差异均无统计学意义($P > 0.05$),具有可比性。

本研究经吉林大学第一医院医学伦理委员会审批通过(25K365-001)。本研究为回顾性研究,仅采集患者既往临床诊疗数据,不干预常规诊疗,经医学伦理委员会审查批准豁免受试者知情同意。

1.2 方法 在实验室信息系统中调取健康对照者体检当日、T1DM 患者入院后首次实验室检查数据,共纳入以下 39 项检测指标。(1)常规生化及免疫指标:包括肌酸激酶(CK)、乳酸脱氢酶(LDH)、天冬氨酸氨基转移酶(AST)、 γ -谷氨酰转肽酶(γ -GGT)、碱性磷酸酶(ALP)、胆碱酯酶(CHE)、淀粉酶(AMY)、脂肪酶(LPS)、总蛋白(TP)、白蛋白(ALB)、球蛋白(GLB)、前白蛋白(PAB)、视黄醇结合蛋白(RBP)、胱抑素 C(Cys C)、总胆红素(TBil)、直接胆红素(DBil)、间接胆红素(IBil)、尿素(Ure)、肌酐(Cre)、胆固醇(CHO)、甘油三酯(TG)、高密度脂蛋白(HDL)、低密度脂蛋白(LDL)、载脂蛋白 A(Apo A)及脂蛋白(a)[Lp(a)]。采用 Roche Cobas 8000 全自动生化免疫分析系统进行常规生化及免疫指标检测,配套试剂购自罗氏诊断产品(上海)有限公司。(2)血常规指标:包括淋巴细胞比率(LR)、血红蛋白(Hb)及血小板计数(PLT)。采用 Sysmex XN-9000 全自动血液分析仪进行血常规检测。(3)铁代谢指标:包括总铁结合能力(TIBC)、铁蛋白(SF)、不饱和铁结合能力(UIBC)

及铁(Fe)。采用 Roche Cobas 8000 系统通过比色法或电化学发光法测定铁代谢指标。(4)微量元素指标:包括镁(Mg)、钙(Ca)、锌(Zn)、铜(Cu)、锌铜比值(Zn/Cu)和硒(Se)。采用 Agilent 7900 电感耦合等离子体质谱仪测定上述元素的血清总浓度。Zn/Cu 由 Zn、Cu 浓度计算得出。(5)糖化血红蛋白(HbA1c):采用 Bio-Rad D-100 全自动糖化血红蛋白分析仪,利用离子交换高效液相色谱法进行检测。采用分层随机抽样法,以分组(T1DM 组/健康对照组)为分层变量,按 7:3 的比例将全部研究对象(T1DM 组 210 例与健康对照组 210 例,共 420 例)划分为训练集($n = 294$)与测试集($n = 126$),训练集与测试集中均同时包含 T1DM 组与健康对照组,且 2 组构成比保持一致。

1.3 统计学处理

1.3.1 统计学分析 采用 Shapiro-Wilk 检验对所有连续变量进行正态性检验,符合正态分布的计量资料以 $\bar{x} \pm s$ 表示,2 组间比较采用独立样本 t 检验;不符合正态分布的计量资料以 $M(P_{25}, P_{75})$ 表示,2 组间比较采用 Mann-Whitney U 检验。所有统计学分析均基于 R 4.2.1 软件实现,检验水准 $\alpha = 0.05$,以 $P < 0.05$ 为差异有统计学意义。

1.3.2 数据预处理与机器学习计算 将 T1DM 组与健康对照组中偏离正常值范围 1.5 倍以上的异常数据进行剔除,并采用随机森林(RF)模型进行缺失值的插补。采用 Lasso 回归结合交叉验证对指标进行二次筛选。该算法通过构造惩罚函数,强制系数绝对值之和小于固定值,并设定不显著因子的回归系数为零,从而自动剔除冗余变量,有效处理高维数据的多重共线性问题。共纳入 8 种机器学习模型,包括极端梯度提升(XGBoost)、Logistic 回归、轻量梯度提升机(LightGBM)、RF、自适应提升(AdaBoost)、决策树(DT)、梯度提升决策树(GBDT)、支持向量机(SVM)。采用 5 折交叉验证方法(其中 4 折作为训练集,1 折作为测试集),整个过程重复 5 次取平均值。

1.3.3 模型性能评估 采用受试者工作特征(ROC)曲线及曲线下面积(AUC)评价模型的区分能力;并在约登指数最大时确定的最佳截断值处构建混淆矩阵,进一步计算准确度、灵敏度、特异度。依据上述指标筛选出最优模型。

1.3.4 MR 计算逻辑 机器学习最终筛选出的实验室指标被视为暴露因素,T1DM 被视为与暴露相关的结果因素。欧洲人群 MR 分析:为了避免样本数据重叠,暴露因素使用来自英国生物银行数据库中与 T1DM 相关的多个独立单核苷酸多态性(SNP,连锁不平衡 < 0.001),这些 SNP 具有全基因组意义($P < 5 \times 10^{-8}$)^[15];对于 T1DM 患者使用来自 9 个队列的 18 942 例 T1DM 病例和 501 638 例欧洲病例对照进行了全基因组关联研究(GWAS)分析^[16]。亚洲人群 MR 分析:暴露因素的遗传工具变量来源于日本生物

银行发布的亚洲人群 GWAS 汇总统计数据,筛选标准与欧洲人群一致($P < 5 \times 10^{-8}$,连锁不平衡 < 0.001)。结局数据采用亚洲人群 T1DM 的 GWAS 汇总数据^[17]。

在 MR 分析中,本研究采用了 4 种互补的方法:逆方差加权法、MR-Egger 回归法、加权中位数法和加权模式法。其中逆方差加权法作为主要分析方法,同时辅以 MR-Egger 回归法、加权中位数法和加权模式法。此外,还使用 MR-Steiger 过滤检验单个 SNP 是更强地与暴露因素相关还是与结局因素相关,从而明确因果关系的方向性。

2 结 果

2.1 训练集与测试集以及训练集中健康对照组与 T1DM 组基线特征、39 项实验室指标比较 训练集与测试集性别、年龄、BMI 及 39 项实验室指标比较,差异均无统计学意义($P > 0.05$),见表 1。训练集中 T1DM 组与健康对照组 HbA1c、CK、LDH、AST、 γ -GGT、ALP、CHE、TP、ALB、GLB、TBil、DBil、IBil、PAB、AMY、LPS、Cre、RBP、Cys C、CHO、TG、HDL、Apo A、Lp(a)、TIBC、SF、UIBC、LR、Hb、PLT、Mg、Ca、Fe 水平比较,差异均有统计学意义($P < 0.05$),见表 2。

表 1 训练集与测试集基线特征及 39 项实验室指标比较[n 或 $M(P_{25}, P_{75})$ 或 $\bar{x} \pm s$]

项目	训练集($n=294$)	测试集($n=126$)	$\chi^2/Z/t$	P
性别			0.200	0.650
男	137	55		
女	157	71		
年龄(岁)	11(8,14)	11(8,13)	0.620	0.535
BMI(kg/m ²)	20.2±3.0	20.3±3.1	-0.310	0.760
HbA1c(%)	5.95(4.90,7.77)	5.75(4.70,8.75)	0.920	0.355
CK(U/L)	95.50(71.25,120.25)	85.00(64.00,108.25)	1.280	0.202
LDH(U/L)	224.00(187.25,263.50)	228.00(188.50,259.25)	0.430	0.665
AST(U/L)	20.95(17.42,26.85)	22.30(17.28,26.02)	0.150	0.884
γ -GGT(U/L)	12.05(9.75,18.35)	11.90(10.07,16.20)	1.380	0.166
ALP(U/L)	121.45(82.00,191.63)	116.50(72.85,193.43)	1.510	0.132
CHE(U/L)	7 721.00(6 983.23,9 102.22)	7 888.35(6 772.60,8 889.77)	1.260	0.207
TP(g/L)	66.49±11.41	68.06±10.86	1.020	0.309
ALB(g/L)	41.33±7.67	42.62±6.78	0.170	0.867
GLB(g/L)	28.95(23.48,33.88)	25.95(22.30,30.65)	0.030	0.973
TBil(μ mol/L)	10.59(9.33,16.91)	12.38(9.10,16.79)	0.720	0.69
DBil(μ mol/L)	3.00(2.05,4.30)	3.60(2.10,5.20)	0.830	0.407
IBil(μ mol/L)	8.80(7.60,13.25)	10.80(7.68,13.93)	0.720	0.451
PAB(g/L)	0.42(0.15,2.20)	0.36(0.16,2.23)	0.500	0.618
AMY(U/L)	64.10(49.85,74.80)	63.45(44.25,82.80)	0.980	0.306
LPS(U/L)	41.85(23.98,61.68)	40.90(26.00,57.45)	0.110	0.914
Ure(mmol/L)	4.46(3.42,5.88)	4.31(3.40,5.30)	0.780	0.435
Cre(μ mol/L)	50.10(38.45,60.65)	49.70(35.88,58.97)	0.880	0.379
RBP(mg/L)	40.15(17.95,57.68)	40.95(19.30,58.78)	0.550	0.582
Cys C(mg/L)	0.74(0.53,0.88)	0.72(0.55,0.94)	0.160	0.869
CHO(mmol/L)	3.92(3.38,4.97)	4.12(3.49,4.67)	1.59	0.114
TG(mmol/L)	1.00(0.68,1.45)	1.02(0.76,1.45)	0.781	0.435
HDL(mmol/L)	1.21(0.97,1.57)	1.25(1.04,1.52)	0.101	0.922
LDL(mmol/L)	2.19(1.66,2.54)	2.21(1.78,2.77)	0.011	0.988
Apo A(g/L)	1.04(0.92,1.27)	1.18(1.02,1.34)	0.420	0.676
Lp(a) (g/L)	0.10(0.04,0.25)	0.14(0.05,0.23)	0.090	0.930
TIBC(μ mol/L)	50.80(36.62,60.90)	47.25(35.42,58.70)	0.520	0.598

续表 1 训练集与测试集基线特征及 39 项实验室指标比较[n 或 $M(P_{25}, P_{75})$ 或 $\bar{x} \pm s$]

项目	训练集($n=294$)	测试集($n=126$)	$\chi^2/Z/t$	P
SF($\mu\text{g/L}$)	100.55(74.28,166.20)	111.80(85.38,202.75)	0.920	0.356
UIBC($\mu\text{mol/L}$)	35.19(27.75,41.01)	32.12(26.23,38.75)	0.520	0.597
LR(%)	0.54(0.42,0.60)	0.53(0.36,0.62)	0.060	0.951
Hb(g/L)	143.00(132.25,150.75)	142.50(132.00,153.00)	1.040	0.302
PLT($\times 10^9/\text{L}$)	255.00(215.25,313.75)	271.50(205.00,303.00)	0.011	0.992
Mg(mmol/L)	9.52(7.38,12.56)	9.92(6.62,14.16)	1.110	0.267
Ca(mmol/L)	59.32 $\pm$ 19.04	61.73 $\pm$ 18.05	0.750	0.455
Fe($\mu\text{mol/L}$)	0.73(0.56,1.04)	0.89(0.61,1.19)	1.640	0.101
Zn($\mu\text{mol/L}$)	0.61(0.44,0.79)	0.60(0.43,0.77)	0.301	0.765
Cu($\mu\text{mol/L}$)	0.63(0.51,0.80)	0.74(0.51,0.93)	0.910	0.361
Zn/Cu	0.82(0.63,1.11)	0.79(0.63,0.97)	0.440	0.659
Se($\mu\text{mol/L}$)	0.10(0.07,0.14)	0.10(0.06,0.14)	0.680	0.496

表 2 训练集中健康对照组与 T1DM 组基线特征及 39 项实验室指标比较[n 或 $M(P_{25}, P_{75})$ 或 $\bar{x} \pm s$]

项目	健康对照组($n=147$)	T1DM 组($n=147$)	$\chi^2/Z/t$	P
性别			0.006	0.812
男	57	60		
女	90	87		
年龄(岁)	11(8,14)	11(9,14)	-0.590	0.558
BMI(kg/m^2)	20.00 $\pm$ 2.90	20.30 $\pm$ 3.30	-0.840	0.404
HbA1c(%)	4.68(4.38,5.06)	8.65(7.24,10.56)	-13.870	<0.001
CK(U/L)	105.51(89.23,125.52)	69.02(57.78,78.72)	11.670	<0.001
LDH(U/L)	240.30(212.91,264.29)	202.96(175.53,231.76)	7.170	<0.001
AST(U/L)	25.29(21.71,28.95)	18.56(16.12,23.05)	8.600	<0.001
γ -GGT(U/L)	12.66(10.39,15.43)	10.42(8.72,13.04)	5.310	<0.001
ALP(U/L)	85.54(71.19,107.21)	195.18(157.73,241.28)	-13.570	<0.001
CHE(U/L)	8 094.61(7 417.64,8 783.55)	7 287.68(6 646.83,8 371.07)	4.300	<0.001
TP(g/L)	75.15 $\pm$ 5.17	61.18 $\pm$ 11.01	13.920	<0.001
ALB(g/L)	46.92 $\pm$ 3.12	37.21 $\pm$ 6.12	17.150	<0.001
GLB(g/L)	29.50(26.73,33.87)	24.04(22.44,27.22)	8.830	<0.001
TBil($\mu\text{mol/L}$)	15.17(12.18,18.90)	10.17(8.07,12.51)	8.840	<0.001
DBil($\mu\text{mol/L}$)	4.80(3.81,5.69)	2.07(1.58,2.72)	13.520	<0.001
IBil($\mu\text{mol/L}$)	13.45(10.94,16.06)	8.26(6.53,10.18)	10.770	<0.001
PAB(g/L)	2.06(1.40,3.10)	0.20(0.15,0.25)	14.810	<0.001
AMY(U/L)	67.87(58.32,77.70)	55.19(45.13,71.13)	5.420	<0.001
LPS(U/L)	54.70(40.72,74.71)	27.46(20.82,39.52)	10.690	<0.001
Ure(mmol/L)	4.40(3.66,5.17)	4.34(3.24,5.53)	0.610	0.542
Cre($\mu\text{mol/L}$)	54.72(46.71,63.00)	35.82(30.48,42.54)	11.860	<0.001
RBP(mg/L)	57.32(51.04,62.47)	18.68(15.80,22.72)	14.800	<0.001
Cys C(mg/L)	0.84(0.68,1.01)	0.69(0.60,0.84)	5.120	<0.001
CHO(mmol/L)	3.94(3.53,4.47)	4.33(3.74,4.89)	-3.780	<0.001
TG(mmol/L)	0.80(0.61,1.02)	1.44(1.03,1.83)	-9.660	<0.001
HDL(mmol/L)	1.30(1.12,1.53)	1.20(1.09,1.37)	3.050	0.002
LDL(mmol/L)	2.13(1.88,2.50)	2.13(1.73,2.58)	0.680	0.498
Apo A(g/L)	1.18(1.07,1.30)	1.01(0.87,1.15)	-7.000	<0.001
Lp(a)(g/L)	0.19(0.14,0.26)	0.10(0.07,0.14)	9.560	<0.001
TIBC($\mu\text{mol/L}$)	61.42(57.95,65.27)	41.36(36.16,48.20)	13.440	<0.001
SF($\mu\text{g/L}$)	182.58(139.43,242.88)	93.38(83.78,111.31)	12.200	<0.001

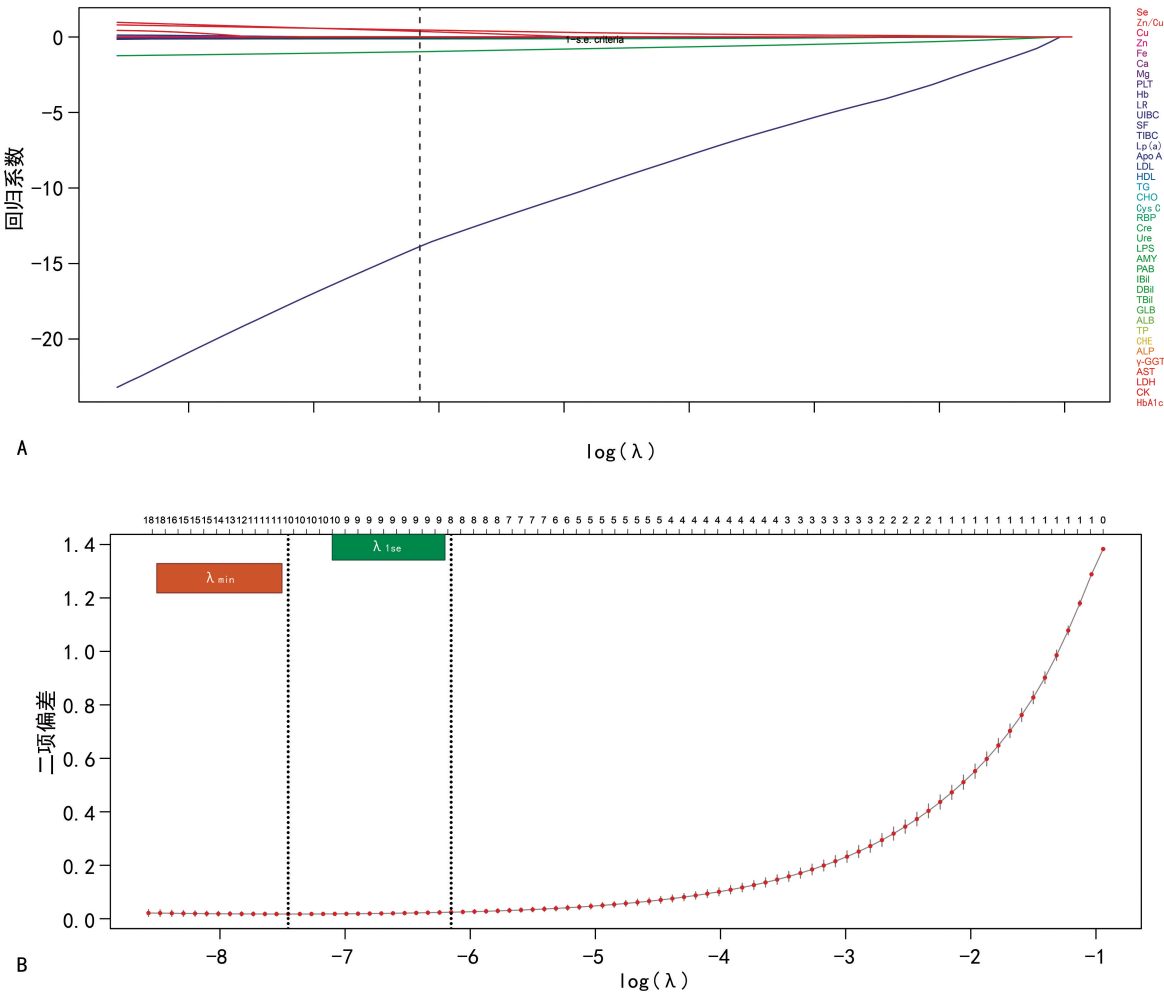
续表 2 训练集中健康对照组与 T1DM 组基线特征及 39 项实验室指标比较[n 或 $M(P_{25}, P_{75})$ 或 $\bar{x} \pm s$]

项目	健康对照组($n=147$)	T1DM 组($n=147$)	$\chi^2/Z/t$	P
UIBC($\mu\text{mol/L}$)	39.06(33.37,44.47)	27.59(22.90,34.31)	9.620	<0.001
LR(%)	0.60(0.54,0.67)	0.41(0.33,0.48)	12.080	<0.001
Hb(g/L)	153.57(143.39,160.77)	133.97(127.67,142.08)	10.770	<0.001
PLT($\times 10^9/\text{L}$)	255.50(216.49,306.75)	274.25(237.94,317.35)	-2.510	0.012
Mg(mmol/L)	12.67(10.38,15.61)	7.35(5.96,9.30)	10.810	<0.001
Ca(mmol/L)	65.41 $\pm$ 18.07	53.67 $\pm$ 15.91	5.910	<0.001
Fe($\mu\text{mol/L}$)	0.96(0.78,1.16)	0.70(0.53,0.92)	6.550	<0.001
Zn($\mu\text{mol/L}$)	0.60(0.50,0.73)	0.59(0.47,0.74)	0.240	0.809
Cu($\mu\text{mol/L}$)	0.73(0.58,0.86)	0.69(0.57,0.84)	1.050	0.295
Zn/Cu	0.81(0.68,0.98)	0.80(0.65,0.93)	0.770	0.444
Se($\mu\text{mol/L}$)	0.10(0.08,0.13)	0.10(0.08,0.14)	-0.470	0.639

2.2 机器学习

2.2.1 特征筛选与变量预处理 本研究纳入 39 个实验室指标作为候选变量。为降低特征维度并减少冗余信息,采用带有 10 折交叉验证的 Lasso 回归进行特征筛选。以交叉验证误差最小对应的 $\lambda_{\min}$ 作为

最优惩罚参数($\lambda_{\min}=0.000\ 578\ 66$,对应 $\log(\lambda)=-7.454\ 8$),最终筛选出 10 个非零系数特征,分别为 CK、AST、TP、ALB、IBil、RBP、SF、TIBC、UIBC 和 Hb,并纳入后续模型构建。见图 1。



注:A为特征变量系数路径图;B为交叉验证曲线图。

图 1 基于 Lasso 回归的 T1DM 特征变量筛选过程

2.2.2 多因素 Logistic 回归分析 以是否为 T1DM (否=0,是=1)为因变量,以 CK、AST、TP、ALB、

IBil、RBP、SF、TIBC、UIBC、Hb 作为自变量(均为原值输入),进行多因素 Logistic 回归分析。结果显示,

RBP、TP、ALB、CK、UIBC、IBil 水平升高均是 T1DM 发生的保护因素($P<0.05$)。见表 3。

表 3 多因素 Logistic 回归分析 T1DM 发生的影响因素

因素	β	SE	Wald χ^2	P	OR	OR 的 95%CI
RBP	-1.842	0.356	26.772	<0.001	0.158	0.079~0.318
TP	-0.181	0.025	51.794	<0.001	0.834	0.794~0.876
ALB	-0.925	0.284	10.608	0.001	0.397	0.227~0.693
CK	-0.542	0.186	8.490	0.004	0.582	0.404~0.838
UIBC	-0.415	0.175	5.624	0.018	0.660	0.468~0.931
IBil	-0.328	0.154	4.536	0.033	0.720	0.532~0.974
TIBC	-0.010	0.007	1.851	0.174	0.990	0.976~1.004
Hb	-0.154	0.125	1.518	0.218	0.857	0.671~1.095
SF	-0.086	0.142	0.367	0.545	0.918	0.695~1.212
AST	-0.042	0.118	0.127	0.722	0.959	0.761~1.208

2.2.3 机器学习模型对比与最优模型选择 将基于 Lasso 回归筛选得到的实验室指标构建 8 种机器学习二分类模型。基于 XGBoost 模型的特征重要性分析, 10 项指标按重要性从高到低依次为 RBP、TIBC、TP、ALB、UIBC、CK、Hb、SF、IBil 和 AST(图 2)。为进一步验证特征筛选结果的稳健性, 本研究同步分析了 RF、LightGBM、GBDT、Logistic 回归、AdaBoost、DT 及 SVM 模型的特征重要性排序(表 4), 结果显示关键变量在不同模型间具有较好一致性。

在训练集中评估 8 种机器学习模型的拟合区分度, 在测试集中绘制 ROC 曲线并计算 AUC 以评价其性能, 见表 5、6。结果显示, XGBoost 模型在测试集中的 AUC 最大(0.968, 95%CI: 0.945~0.991), 且训练集与测试集 AUC 差值最小($\Delta AUC=0.020$), 提示其拟合充分且泛化能力最佳。见图 3。

2.3 MR 分析 对 XGBoost 模型特征重要性排序后的 10 项实验室指标进行双样本 MR 分析, 其效应估计值(OR)及 P 值见表 7。在欧洲人群这 10 项指标的 MR 分析中, 按照以逆方差加权法为主要判定依据、 $P<0.05$ 的标准, 仅 ALB 和 Hb 的逆方差加权法分析结果 $P<0.05$, 存在显著因果关联, 而 TP 的逆方差加权法分析结果 $P>0.05$, 因此不对 TP 做进一步分析。其中 ALB 还同时在加权中位数法和加权模式法分析中 $P<0.05$, 结果更为可靠, 而 Hb 则仅在逆方差加权法分析中 $P<0.05$, 因此对 ALB 进行进一步分析。结果发现, MR-Egger 回归法的截距为 -0.009, $P=0.075$, 表明不存在 SNP 的水平多向性效应。通过异质性统计结果得到 MR-Egger 回归法的 Cochran's Q 值=783.674 6, $P=7.772\ 976\times10^{-72}$; 逆方差加权法的 Cochran's Q 值=796.599 2, $P=$

$1.179\ 697\times10^{-73}$, 结合漏斗图发现 SNP 具有异质性。使用随机效应逆方差加权法(OR=0.900, 95%CI: 0.835~0.970, $P=0.006$)、加权中位数法(OR=0.896, 95%CI: 0.836~0.961, $P=0.002$)和加权模式法(OR=0.900, 95%CI: 0.832~0.974, $P=0.010$)进行分析得到了 186 个能够显著影响 ALB 与 T1DM 因果关系的 SNP。尽管个别 SNP 显示出异质性, 但留一法敏感性分析结果表明, 逐步剔除单个 SNP 后, 逆方差加权法估算的因果效应值保持高度稳定, 且所有估算值的 95%CI 均未跨越零线。这证实了本研究的因果推断结果具有良好的稳健性, 未受到特定离群 SNP 的驱动。除此之外, MR-Steiger 方向性检验结果支持因果方向为 ALB→T1DM, 提示该因果关联的方向正确。而在亚洲人群中, ALB 与 T1DM 的因果关联无统计学意义($P>0.05$)。

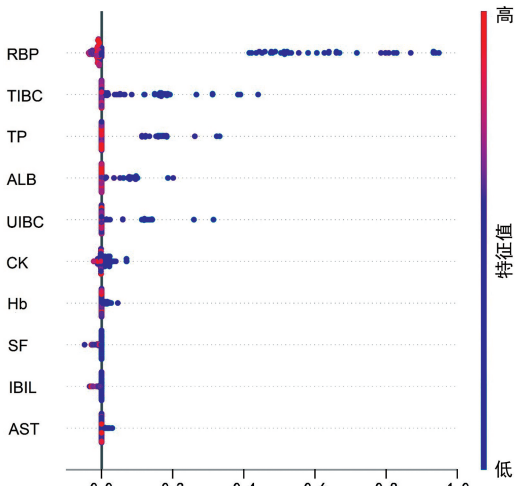


图 2 10 项实验室指标在 XGBoost 模型中的特征重要性排序

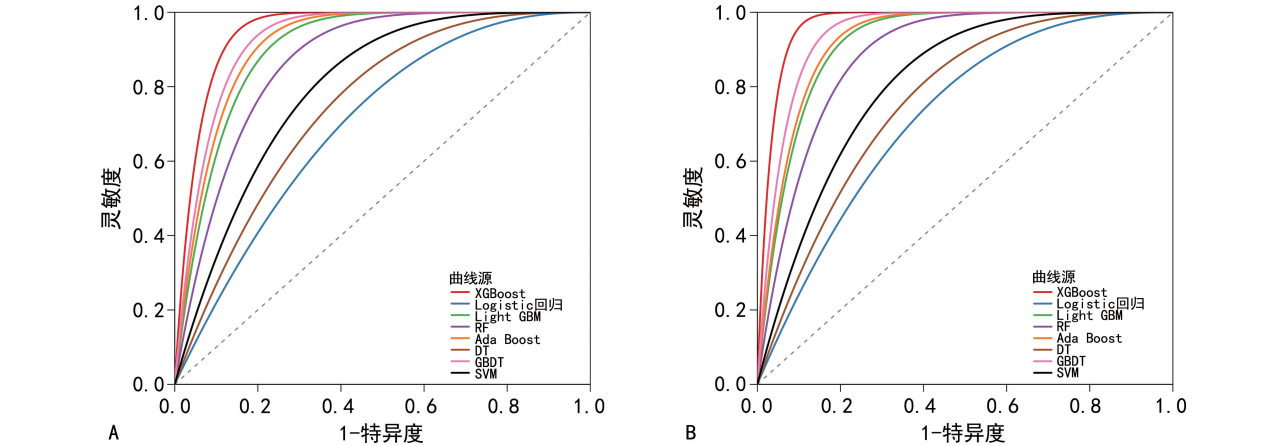
表 4 不同机器学习模型特征重要性排序对比

排名	XGBoost	RF	LightGBM	GBDT	Logistic 回归	AdaBoost	DT	SVM
1	RBP	RBP	RBP	RBP	RBP	RBP	RBP	RBP
2	TIBC	ALB	ALB	ALB	Hb	ALB	ALB	CK

续表 4 不同机器学习模型特征重要性排序对比								
排名	XGBoost	RF	Light GBM	GBDT	Logistic 回归	Ada Boost	DT	SVM
3	TP	CK	CK	CK	CK	CK	TP	ALB
4	ALB	UIBC	UIBC	IBil	SF	UIBC	CK	Hb
5	UIBC	SF	SF	UIBC	TIBC	IBil	UIBC	AST
6	CK	TP	TIBC	TP	ALB	AST	TIBC	TP
7	Hb	IBil	Hb	SF	TP	Hb	Hb	TIBC
8	SF	AST	AST	Hb	UIBC	TIBC	AST	UIBC
9	IBil	Hb	IBil	AST	AST	SF	IBil	SF
10	AST	TIBC	TP	TIBC	IBil	TP	SF	IBil

表 5 8 种机器模型在训练集中的性能							
分类模型	约登指数	<i>P</i>	AUC(95%CI)	准确度	灵敏度	特异度	
XGBoost	0.790	<0.001	0.948 (0.910~0.986)	0.890	0.902	0.888	
Logistic 回归	0.350	<0.001	0.701 (0.620~0.782)	0.680	0.667	0.683	
LightGBM	0.702	<0.001	0.901 (0.850~0.952)	0.850	0.833	0.869	
RF	0.620	<0.001	0.866 (0.804~0.928)	0.820	0.792	0.828	
AdaBoost	0.715	<0.001	0.914 (0.867~0.961)	0.860	0.845	0.870	
DT	0.430	<0.001	0.748 (0.662~0.834)	0.720	0.701	0.730	
GBDT	0.741	<0.001	0.926 (0.889~0.963)	0.875	0.860	0.881	
SVM	0.520	<0.001	0.798 (0.729~0.867)	0.765	0.748	0.772	

表 6 8 种机器模型在测试集中的性能							
分类模型	约登指数	<i>P</i>	AUC(95%CI)	准确度	灵敏度	特异度	
XGBoost	0.842	<0.001	0.968 (0.945~0.991)	0.921	0.937	0.905	
Logistic 回归	0.397	<0.001	0.724 (0.635~0.813)	0.698	0.683	0.714	
Light GBM	0.730	<0.001	0.918 (0.875~0.961)	0.865	0.841	0.889	
RF	0.667	<0.001	0.884 (0.830~0.938)	0.833	0.810	0.857	
Ada Boost	0.746	<0.001	0.925 (0.882~0.968)	0.873	0.857	0.889	
DT	0.476	<0.001	0.765 (0.680~0.850)	0.738	0.714	0.762	
GBDT	0.778	<0.001	0.942 (0.910~0.974)	0.889	0.873	0.905	
SVM	0.556	<0.001	0.812 (0.745~0.879)	0.778	0.762	0.794	



注:A 为训练集不同机器学习模型 ROC 曲线对比图;B 为测试集不同机器学习模型 ROC 曲线对比图。

图 3 不同机器学习模型 ROC 曲线对比图

表 7 基于机器学习模型筛选的实验室指标 MR 分析统计值

方法	统计值	亚洲人群									
		RBP	TIBC	TP	ALB	UIBC	CK	Hb	SF	IBil	AST
MR-Egger 回归法	<i>P</i>	—	—	0.742	0.451	—	0.959	0.527	—	—	0.008
	<i>OR</i>	—	—	1.167	0.477	—	0.777	0.913	—	—	7.183
加权中位数法	<i>P</i>	—	—	0.532	0.103	—	0.650	0.817	—	—	0.279
	<i>OR</i>	—	—	0.836	0.503	—	1.012	1.016	—	—	1.373
逆方差加权法	<i>P</i>	—	—	0.427	0.375	—	0.295	0.823	—	—	0.357
	<i>OR</i>	—	—	0.861	1.366	—	1.142	1.002	—	—	1.326
加权模式法	<i>P</i>	—	—	0.456	0.088	—	0.654	0.942	—	—	0.519
	<i>OR</i>	—	—	0.798	0.422	—	1.175	1.005	—	—	1.302

方法	统计值	欧洲人群									
		RBP	TIBC	TP	ALB	UIBC	CK	Hb	SF	IBil	AST
MR-Egger 回归法	<i>P</i>	—	—	0.046	0.965	—	0.263	0.098	0.311	—	0.272
	<i>OR</i>	—	—	0.906	1.007	—	0.886	1.348	1.290	—	1.302
加权中位数法	<i>P</i>	—	—	0.033	0.002	—	0.989	0.563	0.414	—	0.140
	<i>OR</i>	—	—	1.051	0.896	—	1.022	1.061	1.100	—	1.293
逆方差加权法	<i>P</i>	—	—	0.662	0.007	—	0.693	0.008	0.444	—	0.217
	<i>OR</i>	—	—	1.016	0.900	—	0.989	1.377	1.080	—	1.196
加权模式法	<i>P</i>	—	—	0.697	0.018	—	0.492	0.565	0.260	—	0.023
	<i>OR</i>	—	—	1.018	0.900	—	1.076	1.071	1.208	—	1.247

注：—表示数据不可用。

3 讨 论

3.1 机器学习与 MR 整合策略的方法学价值 本研究证实,将机器学习的高灵敏度筛选与 MR 的因果推断相结合,是从高维临床数据中识别疾病风险因子的有效策略。通过这一整合分析框架,本研究不仅从 39 项常规入院检测指标中精准定位了与 T1DM 具有潜在因果关联的保护因素 ALB,还利用 MR 分析有效剔除了一些虽然在临床上显著相关但缺乏遗传因果证据的生物标志物(如 CK、AST)。这一结果表明,该整合策略能够克服传统观察性研究易受混杂因素干扰的局限性,为利用现有的临床存量数据进行疾病早期预警和发病机制探索提供了可靠的方法学支撑。

本研究在机器学习阶段通过 Lasso 回归从 39 项实验室指标中筛选出与 T1DM 显著关联的候选变量后,进一步采用双样本 MR 方法,主要目的是从关联性层面提升到因果推断层面。具体而言,机器学习筛选出的实验室指标(如特征重要性排名前 10 的变量)仅能显示统计学上的相关关系,但无法排除混杂因素、反向因果或双向关联的干扰。MR 分析利用遗传变异(SNP)作为工具变量,基于孟德尔遗传定律的随机分配原理,天然满足工具变量的 3 个核心假设(相关性、独立性、排他性),从而能在很大程度上规避传统观察性研究的混杂偏倚和反向因果问题。通过

MR-Steiger 过滤等技术,还能明确因果方向,即判断是实验室指标影响 T1DM 发生,还是 T1DM 导致实验室指标改变。本研究分别在欧洲人群和亚洲人群中进行跨种族验证,既能检验候选指标的因果关联是否具有人群一致性,又能增强结果的稳健性和普适性,为后续功能验证或临床转化提供更可靠的生物学证据。

在方法学层面,本研究的“机器学习筛选—MR 验证”两步策略与近年来兴起的数据驱动因果推断范式一致。Yuan 等^[17]通过广角 MR 分析系统评估了多种代谢因素与 T1DM 的因果关系,证实了 MR 在筛选 T1DM 风险因素中的独特价值。然而,该研究的暴露因素选择基于先验知识,而非数据驱动。本研究的创新之处在于利用机器学习从高维数据中无偏地生成候选假说,随后通过 MR 进行因果检验,形成了“假说生成—假说检验”的完整闭环,有效降低了传统研究中因选择性报告导致的偏倚风险。此外,本研究的 8 种机器学习模型比较显示,XGBoost 模型表现最优(AUC=0.968),这与 Ogunleye 等^[18]在慢性肾脏病预测中报道的 XGBoost 模型优越性一致,进一步证实了集成学习模型在处理中等规模、高维度临床数据时的优势。

3.2 ALB 水平升高作为 T1DM 保护因素的因果证

据与机制解析 本研究的核心发现之一是 ALB 在 T1DM 发病机制中的潜在因果作用。机器学习分析显示 ALB 的特征重要性排名均靠前列,表明该指标具有良好的稳健性;同时 T1DM 组 ALB 水平显著低于对照组;多因素 Logistic 回归分析进一步证实 ALB 水平升高为 T1DM 发生的独立保护因素($OR = 0.397, P = 0.001$)。MR 分析在欧洲人群中证实,遗传预测的 ALB 水平降低与 T1DM 风险增加存在因果关联(随机效应逆方差加权法: $OR = 0.900, P = 0.006$)。

在横向比较方面,Kunutsor 等^[19]在芬兰开展的前瞻性队列研究中观察到血清 ALB 水平每升高 1 g/L,2 型糖尿病(T2DM)风险降低约 5%,提示 ALB 的保护效应可能并不局限于某一种糖尿病亚型。然而,T1DM 与 T2DM 的发病机制存在本质差异——前者以自身免疫介导的胰岛 β 细胞破坏为核心,后者以胰岛素抵抗为主。因此,ALB 对 T1DM 的保护机制可能涉及不同于 T2DM 的特异性通路。

从分子机制角度分析发现 ALB 对 T1DM 的保护作用可能通过以下途径实现。首先,ALB 是血浆中主要的抗氧化剂,含有特定的巯基结构(Cys-34 位点),能够清除自由基并抑制脂质过氧化^[20]。Belinskaia 等^[21]指出,ALB 的抗氧化活性占血浆总抗氧化能力的 70% 以上,且其氧化修饰形式(缺血修饰白蛋白、高级氧化蛋白产物)可作为氧化应激的敏感标志物。T1DM 的发病伴随着胰岛 β 细胞的氧化应激损伤,低水平的 ALB 可能导致机体抗氧化能力减弱,从而加速胰岛 β 细胞的凋亡与功能衰竭^[22]。其次,ALB 是一种负性急性时相反应蛋白,其水平下降往往提示体内存在慢性低度炎症状态。Soeters 等^[23]研究表明,低白蛋白血症的核心机制并非营养不良,而是炎症驱动的毛细血管通透性增加和分解代谢增强。在 T1DM 的自身免疫进程中,持续的低度炎症不仅直接攻击胰岛 β 细胞,还可能通过降低 ALB 水平削弱机体的抗氧化防御,形成“炎症—低 ALB—氧化应激—胰岛 β 细胞损伤”的恶性循环^[24]。

3.3 跨种族 MR 分析结果差异与解释 值得注意的是,本研究的 MR 分析显示 ALB 与 T1DM 的因果关联仅在欧洲人群中显著($P < 0.05$),而在亚洲人群中关联无统计学意义($P > 0.05$)。本研究机器学习阶段基于吉林人群临床数据,MR 验证阶段分别在欧洲和亚洲人群 GWAS 数据中进行,这种“跨种族发现—验证”设计在 MR 研究中具有方法学依据。ALB 作为高度保守的血浆蛋白,其抗氧化和免疫调节功能不受种族限制^[25-26],因此机器学习阶段在我国人群中筛选出的 ALB-T1DM 关联假说具有跨种族验证的生物学

基础。Porcu 等^[27]研究表明,跨种族 MR 分析可以增强因果推断的外部效度,同时有助于区分种族特异性的遗传效应。

亚洲人群中未检测到显著因果关联,可能的原因包括:(1)统计效能不足——亚洲人群 T1DM 的 GWAS 样本量远小于欧洲人群,导致 MR 分析检测微弱因果效应的能力受限;(2)遗传结构差异——Kawabata 等^[28]研究表明,T1DM 在亚洲人群中具有独特的遗传结构,人类白细胞抗原(HLA)区域的易感位点与欧洲人群存在差异,调节 ALB 水平的关键 SNP 在 2 个种族中的等位基因频率和连锁不平衡结构也可能不同;(3)基因-环境交互作用——亚洲人群的饮食结构(蛋白质摄入量和来源)和生活方式可能修饰了遗传变异对 ALB 水平的效应,从而弱化了可检测到的因果信号。

3.4 其他关键指标的临床价值与 MR 过滤效应 除 ALB 外,本研究还揭示了其他指标的临床价值及 MR 的“过滤”作用。RBP 在多种机器学习模型中重要性均排名第 1,且其水平在 T1DM 患者中显著降低。Christou 等^[29]指出,RBP4 不仅是反映营养状态及视黄醇代谢的关键蛋白,还通过调节胰岛素信号通路和脂肪组织功能参与代谢调控。Forga 等^[30]在 T1DM 成人患者中观察到 RBP 水平显著降低,与本研究机器学习模型筛选结果一致。遗憾的是,受限于现有 GWAS 数据库中 RBP 相关遗传工具变量的可用性,本研究未能完成 RBP 的 MR 验证,但这为后续研究指明了方向。

相比之下,CK、AST 以及铁代谢相关指标(Hb、TIBC、UIBC)虽在组间存在差异,且既往文献提示铁过载或与胰岛素抵抗及胰岛损伤相关,但本研究的 MR 分析未支持它们与 T1DM 的直接因果关联^[31]。Jin 等^[32]总结了铁代谢与糖尿病的多层面关联,指出铁过载主要通过氧化应激和胰岛素抵抗途径影响 T2DM,而对 T1DM 的直接因果作用证据有限。Vasudevan 等^[33]通过双向 MR 分析也未发现铁代谢指标与 T1DM 之间存在稳健的因果关联,这与本研究的 MR 分析结果高度一致。这提示这些指标的改变更可能是 T1DM 代谢紊乱导致的继发性结果(如高血糖导致肌肉损伤或炎症引起的铁代谢异常),而非致病原因。这一发现再次凸显了引入 MR 分析剔除“假阳性”因果的必要性。

3.5 39 项实验室指标纳入的合理性与临床转化价值 本研究纳入了涵盖 39 项实验室指标的检测项目。从临床实用性角度看,虽然在常规诊疗中同时检测 39 项指标的成本较高,但在大数据驱动的临床研究语境下,这种广谱筛选具有显著优势:(1)T1DM 不

仅是糖代谢异常,更涉及复杂的微环境紊乱,如微量元素平衡与铁代谢异常。纳入 Zn、Se、PAB 等指标,有助于从多维度捕捉 T1DM 的病理生理特征。(2)本研究通过机器学习模型对 39 项指标进行筛选,通过大规模数据分析验证,最终识别出 ALB 等少数关键变量。(3)这种基于全项检查后筛选出的核心模型,相比于传统的经验性选择指标,更具科学性和稳健性,为临床制订精准的个性化诊疗方案提供了新思路。

在临床转化方面,虽然 MR 结果提示 ALB 水平升高可能对 T1DM 具有保护作用,但在实际应用中发现 ALB 水平极易受到营养状况、肝肾功能及急性感染等多种因素的影响,且波动范围较大。因此,单纯依靠一次 ALB 检测值进行风险评估并不准确。笔者建议在临床实践中,应将 ALB 视为一个长期的监测指标,结合 C 反应蛋白等炎症标志物进行综合评估。对于 T1DM 高危人群,通过合理的营养干预维持正常的 ALB 水平,可能有助于增强机体抗氧化能力,延缓疾病进展。

3.6 研究局限性、挑战与未来方向 本研究存在以下局限性需要说明。(1)样本量与代表性方面:本研究纳入的 420 例研究对象均来自单一中心,样本量相对有限,可能存在选择偏倚。机器学习模型的泛化能力仍需在多中心、大样本队列中进一步验证。(2)指标覆盖范围方面:虽然本研究纳入了 39 项实验室指标,覆盖了肝肾功能、血脂、血常规、铁代谢及微量元素等多个维度,但可能遗漏了具有更高特异性的免疫学标志物,如胰岛细胞抗体(ICA)、谷氨酸脱羧酶抗体(GAD-Ab)及胰岛素自身抗体(IAA)等 T1DM 特异性自身抗体。纳入这些指标可能进一步提升模型的评估性能和生物学解释力。(3)MR 分析的局限性:本研究 MR 分析依赖于公共 GWAS 汇总数据,暴露与结局数据来源于不同队列,可能存在样本人群结构不完全匹配的问题。此外,尽管 MR-Egger 截距检验未检出显著的水平多向性($P=0.075$),但 Cochran's Q 检验提示存在异质性,这可能源于多效性 SNP 的残留影响。(4)跨种族外推性方面:如前所述,亚洲人群 T1DM 的 GWAS 数据样本量有限,本研究在亚洲人群中的 MR 结论需持谨慎态度,不宜直接外推至我国人群。

面对上述局限性,未来研究可从以下方向加以改进和深化。(1)扩大样本与外部验证:联合国内多家医疗中心开展多中心研究,纳入更大规模的 T1DM 队列进行机器学习模型的外部验证,同时覆盖不同地区和民族的 T1DM 患者,以提高研究结论的代表性和普适性。(2)整合多组学数据:结合蛋白质组学、代谢组学和转录组学数据,从多维度解析 ALB 等关键分子

在 T1DM 发病中的具体分子通路,例如通过 pQTL-MR 分析直接利用蛋白质水平的遗传工具变量进行因果推断,可望提高 MR 分析的精度和生物学可解释性。(3)利用新兴亚洲人群数据库:随着亚洲人群大型队列的 GWAS 数据持续积累,未来有望获得更大样本量的亚洲 T1DM 遗传数据,从而在亚洲人群中重新检验 ALB 与 T1DM 的因果关联。(4)前瞻性临床干预研究:基于本研究发现 ALB 水平升高为 T1DM 保护因素的线索,可设计前瞻性队列研究或干预试验,观察通过营养支持维持正常 ALB 水平是否能延缓 T1DM 高危人群的疾病进展,为临床转化提供直接证据。

综上所述,本研究成功构建并验证了整合机器学习筛选与 MR 验证的分析框架,从 39 项常规实验室指标中识别出 ALB 水平升高为 T1DM 的潜在因果性保护因素。该策略为临床大数据驱动的因果推断提供了可推广的方法学范式,同时为 T1DM 的风险评估与早期干预提供了新的生物标志物线索。

利益冲突 所有作者均声明不存在利益冲突。

作者贡献 李贻:研究构思,论文撰写与修改;杜佳:资料收集整理;田肃岩:研究思路制订,方法学指导;关贺元:数据分析;孙琪:数据分析;许建成:研究思路制订,论文修改。

参考文献

- [1] Mobasser M, Shirmohammadi M, Amiri T, et al. Prevalence and incidence of type 1 diabetes in the world: a systematic review and Meta-analysis[J]. Health Promot Perspect, 2020, 10(2): 98-115.
- [2] Gregory G A, Robinson T I G, Linklater S E, et al. Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study[J]. Lancet Diabetes Endocrinol, 2022, 10(10): 741-760.
- [3] Ogle G D, James S, Dabelea D, et al. Global estimates of incidence of type 1 diabetes in children and adolescents: results from the International Diabetes Federation Atlas, 10th edition [J]. Diabetes Res Clin Pract, 2022, 183: 109083.
- [4] DiMeglio L A, Evans-Molina C, Oram R A. Type 1 diabetes[J]. Lancet, 2018, 391(10138): 2449-2462.
- [5] Lombardo F, Passanisi S, Gasbarro A, et al. Hepatomegaly and type 1 diabetes: a clinical case of Mauriac's syndrome[J]. Ital J Pediatr, 2019, 45(1): 3.
- [6] Heinrich N S, Theilade S, Winther S A, et al. Copeptin and renal function decline, cardiovascular events and mortality in type 1 diabetes [J]. Nephrol Dial Transplant, 2021, 37(1): 100-107.
- [7] Wolnik B, Orłowska-Kunikowska E, Błaszowska M, et

- al. The phenomenon of HbA1c stability and the risk of hypoglycemia in long-standing type 1 diabetes[J]. *Diabetes Res Clin Pract*, 2019, 152: 96-102.
- [8] Oost L J, van Heck J I P, Tack C J, et al. The association between hypomagnesemia and poor glycaemic control in type 1 diabetes is limited to insulin resistant individuals [J]. *Sci Rep*, 2022, 12(1): 6433.
- [9] Davey Smith G, Hemani G. Mendelian randomization; genetic anchors for causal inference in epidemiological studies[J]. *Hum Mol Genet*, 2014, 23(R1): R89-R98.
- [10] Richmond R C, Davey Smith G. Mendelian randomization: concepts and scope[J]. *Cold Spring Harb Perspect Med*, 2022, 12(1): a040501.
- [11] Qin H, Yang K, Wang H, et al. A causal relationship between type 1 diabetes and risk of osteoporosis: a univariable and multivariable Mendelian randomization study [J]. *J Diabetes Res*, 2024, 2024: 1610688.
- [12] Zhao S, Li Y, Su C. Assessment of common risk factors of diabetes and chronic kidney disease: a Mendelian randomization study [J]. *Front Endocrinol*, 2023, 14: 1265719.
- [13] Liu N, Wang G, Liu C, et al. Non-alcoholic fatty liver disease and complications in type 1 and type 2 diabetes: a Mendelian randomization study [J]. *Diabetes Obes Metab*, 2023, 25(2): 365-376.
- [14] Chen Z, Boehnke M, Wen X, et al. Revisiting the genome-wide significance threshold for common variant GWAS [J]. *G3 Genes Genomes Genetics*, 2021, 11(2): jkaa056.
- [15] Chiou J, Geusz R J, Okino M L, et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics [J]. *Nature*, 2021, 594(7863): 398-402.
- [16] Kanai M, Akiyama M, Takahashi A, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases [J]. *Nat Genet*, 2018, 50(3): 390-400.
- [17] Yuan S, Larsson S C. An atlas on risk factors for type 2 diabetes: a wide-angled Mendelian randomisation study [J]. *Diabetologia*, 2020, 63(11): 2359-2371.
- [18] Ogunleye A, Wang Q G. XGBoost model for chronic kidney disease diagnosis[J]. *IEEE/ACM Trans Comput Biol Bioinform*, 2020, 17(6): 2131-2140.
- [19] Kunutsor S K, Khan H, Laukkanen J A. Serum albumin concentration and incident type 2 diabetes risk; new findings from a population-based cohort study[J]. *Diabetologia*, 2015, 58(5): 961-967.
- [20] Zhang Y, Jiang R, Shi Z, et al. Oxidized albumin induces renal tubular cell death and promotes the progression of renal diseases through ferroptosis [J]. *Int J Mol Sci*, 2025, 26(13): 5924.
- [21] Belinskaia D A, Voronina P A, Shmurak V I, et al. Serum albumin in health and disease: esterase, antioxidant, transporting and signaling properties [J]. *Int J Mol Sci*, 2021, 22(19): 10318.
- [22] Herold K C, Delong T, Perdigo A L, et al. The immunology of type 1 diabetes [J]. *Nat Rev Immunol*, 2024, 24(6): 435-451.
- [23] Soeters P B, Wolfe R R, Shenkin A. Hypoalbuminemia; pathogenesis and clinical significance [J]. *J Parenter Enter Nutr*, 2019, 43(2): 181-193.
- [24] Peng J, Xiang J, Chen Z, et al. CALLY index and hypertension in diabetes: a composite inflammation-nutrition-immunity marker [J]. *Sci Prog*, 2025, 108(4): 368504251387822.
- [25] Ward E S, Gelinas D, Dreesen E, et al. Clinical significance of serum albumin and implications of FeRn inhibitor treatment in IgG-mediated autoimmune disorders [J]. *Front Immunol*, 2022, 13: 892534.
- [26] Guo J, Shi Y, Wang Y, et al. Studying of anti-inflammatory and antioxidant effects of tectorigenin in ovalbumin-induced asthma mice models [J]. *Clin Respir J*, 2024, 18(4): e13742.
- [27] Porcu E, Sadler M C, Lepik K, et al. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome [J]. *Nat Commun*, 2021, 12: 5647.
- [28] Kawabata Y, Ikegami H. Genetics of fulminant type 1 diabetes [J]. *Diabetol Int*, 2020, 11(4): 315-322.
- [29] Christou G, Tselepis A, Kiortsis D. The metabolic role of retinol binding protein 4: an update [J]. *Horm Metab Res*, 2012, 44(1): 6-14.
- [30] Forga L, Bolado F, Goñi M J, et al. Low serum levels of prealbumin, retinol binding protein, and retinol are frequent in adult type 1 diabetic patients [J]. *J Diabetes Res*, 2016, 2016: 2532108.
- [31] González-Villalva A, Colín-Barenque L, Bizarro-Nevarés P, et al. Pollution by metals: is there a relationship in glycemic control? [J]. *Environ Toxicol Pharmacol*, 2016, 46: 337-343.
- [32] Jin E J, Jo Y, Wei S, et al. Ferroptosis and iron metabolism in diabetes: pathogenesis, associated complications, and therapeutic implications [J]. *Front Endocrinol*, 2024, 15: 1447148.
- [33] Vasudevan A, Venkatesan P. Association of organ iron levels with type 2 diabetes mellitus and glycemic traits: a bidirectional two-sample Mendelian randomization study [J]. *J Trace Elem Med Biol*, 2025, 87: 127586.