

利用膳食抗氧化剂识别心血管疾病人群合并慢性肾病：基于机器学习与SHAP分析

朱 颖

福建师范大学数学与统计学院, 福建 福州

收稿日期: 2026年8月15日; 录用日期: 2026年9月7日; 发布日期: 2026年9月17日

摘 要

背景与目的: 慢性肾病(CKD)是心血管疾病(CVD)人群中的常见共病, 并与不良预后密切相关。然而, CVD人群中合并CKD的关键识别特征尚未充分明确, 膳食抗氧化剂在疾病识别中的作用也缺乏系统评估。本研究旨在采用可解释机器学习方法构建CVD人群CKD识别模型, 并分析膳食抗氧化剂及其他相关特征对模型输出的贡献。**方法:** 本研究基于美国国家健康与营养调查(NHANES) 2007~2008年、2009~2010年及2017~2018年数据, 纳入年龄 ≥ 20 岁的CVD参与者, 排除妊娠、膳食资料不可靠、能量摄入异常及关键变量缺失者。CKD判定标准为估算肾小球滤过率 $< 60 \text{ mL/min/1.73 m}^2$ 或尿白蛋白肌酐比 $\geq 30 \text{ mg/g}$ 。采用逻辑回归、随机森林、极端随机树、梯度提升树、XGBoost和LightGBM构建CKD识别模型, 并以AUC-ROC、AUC-PR等指标评价模型性能。进一步采用SHAP方法解释膳食抗氧化剂及其他关键特征对模型输出的贡献。**结果:** 共纳入946名CVD参与者, 其中单纯CVD组560人, CVD合并CKD组386人。XGBoost模型的区分能力最佳, AUC-ROC为0.852, AUC-PR为0.726; 在默认阈值0.50下, 其准确率、敏感度和特异度分别为0.742、0.781和0.723。SHAP分析显示, 年龄、中高强度体力活动、糖尿病、高血压及体质指数等基础特征对模型输出贡献较大。膳食抗氧化剂中, 芹菜素、矢车菊素、槲皮素、维生素E、镁、染料木素和柚皮素等的SHAP重要性相对较高; 部分连续特征与模型输出之间呈现非线性关系。**结论:** 在CVD人群中, 除年龄、体力活动及代谢性疾病等基础特征外, 部分膳食抗氧化剂也可同期CKD状态识别提供补充信息。基于XGBoost的可解释机器学习模型具有较好的区分能力, 但本研究为横断面分析, 结果仅用于识别参与者当前是否合并CKD, 不能用于预测未来CKD发生。未来仍需通过纵向研究和外部样本进一步验证。

关键词

心血管疾病, 慢性肾病, 膳食抗氧化剂, 机器学习, SHAP分析

Identification of Coexisting Chronic Kidney Disease among Individuals with Cardiovascular Disease Using Dietary Antioxidants: A Machine Learning and SHAP Analysis

Ying Zhu

School of Mathematics and Statistics, Fujian Normal University, Fuzhou Fujian

Received: August 15, 2026; accepted: September 7, 2026; published: September 17, 2026

Abstract

Background and Objective: Chronic kidney disease (CKD) is a common comorbidity among individuals with cardiovascular disease (CVD) and is closely associated with adverse outcomes. However, the key features for identifying coexisting CKD in the CVD population have not been fully clarified, and the role of dietary antioxidants in disease identification has not been systematically evaluated. This study aimed to develop an interpretable machine learning model for identifying CKD among individuals with CVD and to assess the contributions of dietary antioxidants and other relevant features to model output. **Methods:** This study used data from the 2007~2008, 2009~2010, and 2017~2018 cycles of the National Health and Nutrition Examination Survey (NHANES). Participants aged ≥ 20 years with CVD were included. Those who were pregnant or had unreliable dietary data, implausible energy intake, or missing key variables were excluded. CKD was defined as an estimated glomerular filtration rate < 60 mL/min/1.73 m² or a urinary albumin-to-creatinine ratio ≥ 30 mg/g. Logistic regression, random forest, extremely randomized trees, gradient boosting, XGBoost, and LightGBM were used to develop models for identifying CKD. Model performance was evaluated using the area under the receiver operating characteristic curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), and other performance metrics. SHapley Additive exPlanations (SHAP) was further used to interpret the contributions of dietary antioxidants and other key features to model output. **Results:** A total of 946 participants with CVD were included, of whom 560 had CVD alone and 386 had coexisting CKD. XGBoost showed the best discrimination, with an AUC-ROC of 0.852 and an AUC-PR of 0.726. At the default threshold of 0.50, its accuracy, sensitivity, and specificity were 0.742, 0.781, and 0.723, respectively. SHAP analysis showed that age, moderate-to-vigorous physical activity, diabetes, hypertension, and body mass index made relatively large contributions to model output. Among dietary antioxidants, apigenin, cyanidin, quercetin, vitamin E, magnesium, genistein, and naringenin showed relatively high SHAP importance. Several continuous features also showed nonlinear relationships with model output. **Conclusion:** Among individuals with CVD, several dietary antioxidants may provide supplementary information for identifying concurrent CKD, in addition to basic features such as age, physical activity, and metabolic conditions. The interpretable XGBoost model showed good discriminatory ability. However, because this was a cross-sectional analysis, the model is intended only to identify current CKD status and cannot be used to predict future CKD onset. Further validation in longitudinal studies and external samples is needed.

Keywords

Cardiovascular Disease, Chronic Kidney Disease, Dietary Antioxidants, Machine Learning,

SHAP Analysis

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

心血管疾病(cardiovascular disease, CVD)与慢性肾病(chronic kidney disease, CKD)常常相伴发生。高血压、糖尿病等代谢异常与两类疾病的发生发展密切相关[1], 动脉粥样硬化、炎症反应及神经体液激活也参与心肾相互损伤[2]。在已有 CVD 的人群中, 估算肾小球滤过率下降或出现白蛋白尿可提示肾脏受累, 并与较高的肾脏及心血管不良结局风险相关[3]。因此, 识别 CVD 人群中合并 CKD 的个体, 对肾脏评估和风险分层具有现实意义。

氧化应激可能是心血管损伤与肾功能异常之间的共同环节。活性氧生成增加而抗氧化防御不足时, 可导致内皮功能障碍、脂质过氧化增强和炎症反应加重, 进而累及血管和肾脏组织[4]。膳食中的抗氧化维生素、镁、硒、类胡萝卜素及黄酮类成分可参与机体抗氧化防御。基于 NHANES 数据的研究显示, 较高的复合膳食抗氧化指数与较低的 CKD 和 CVD 患病可能性相关[5][6]; 但一项 CKD 队列研究发现, 氧化平衡评分与终末期肾病及心血管结局的关联在充分调整后有所减弱[7]。膳食抗氧化相关指标在不同人群中的意义仍需进一步评价。

已有研究将膳食抗氧化剂与人口学、生活方式及健康状况特征共同纳入机器学习模型, 用于识别 CVD 与癌症共病以及糖尿病人群中的 CVD 状态, 并通过 SHapley Additive exPlanations (SHAP)方法解释模型输出[8][9]。可解释机器学习也已用于 CKD 识别和进展风险评估[10][11]。不过, 针对 CVD 人群是否合并 CKD, 并同时评价多种具体膳食抗氧化剂贡献的研究仍较少。此外, 多种抗氧化剂摄入量之间可能存在较强相关性, 单独考察某一种营养素难以反映其与年龄、体力活动及心肾代谢疾病共同形成的特征结构。

基于此, 本研究利用 NHANES 2007~2008 年、2009~2010 年及 2017~2018 年数据, 以 CVD 参与者是否合并 CKD 为识别目标, 同时纳入膳食抗氧化剂、人口学、生活方式及健康状况特征, 比较逻辑回归、随机森林、极端随机树、梯度提升树、XGBoost 和 LightGBM 六种模型的识别性能, 并采用 SHAP 方法分析最优模型的特征贡献及潜在非线性关系。本研究旨在识别区分单纯 CVD 与 CVD 合并 CKD 的重要特征, 评价膳食抗氧化剂的补充识别价值, 为 CVD 人群的肾脏评估和风险分层提供参考。

2. 方法

2.1. 参与者

美国国家健康与营养调查(NHANES)由美国疾病控制与预防中心下属的国家健康统计中心组织实施, 采用分层、多阶段概率抽样方法, 通过访谈、体检及实验室检测收集具有全国代表性的健康与营养数据。

本研究合并分析 NHANES 2007~2008 年、2009~2010 年及 2017~2018 年三个调查周期的数据, 纳入年龄 ≥ 20 岁的成年参与者。排除标准包括: 妊娠期女性; 未完成两次可靠 24 小时膳食回顾者; 能量摄入异常者(男性 < 800 或 > 4200 kcal/天, 女性 < 600 或 > 3500 kcal/天); 以及 CVD 诊断信息、CKD 判定相关实验室指标、膳食抗氧化剂摄入或其他关键变量缺失者。随后进一步筛选 CVD 参与者作为研究对象,

并依据 CKD 状态分为单纯 CVD 组和 CVD 合并 CKD 组。最终共纳入 946 名 CVD 参与者, 其中单纯 CVD 组 560 人, CVD 合并 CKD 组 386 人。详细的参与者筛选流程见图 1。

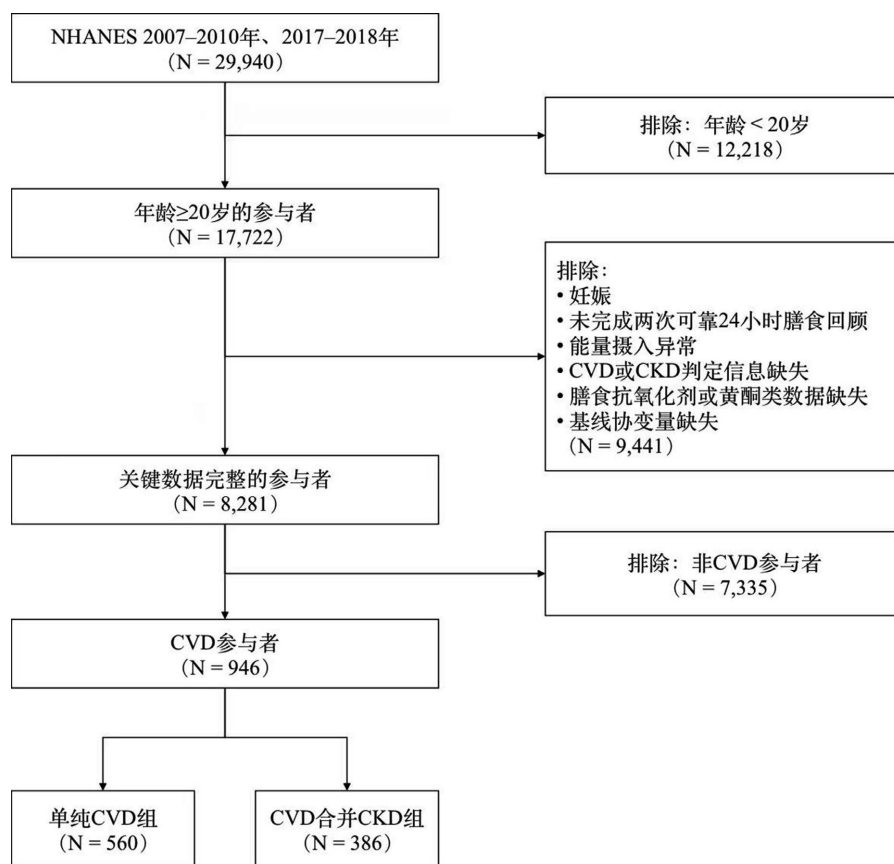


Figure 1. Participant selection process

图 1. 参与者筛选流程

2.2. 膳食抗氧化剂暴露评估

本研究从 NHANES 数据库中提取了 44 种膳食抗氧化剂的摄入量数据, 涉及维生素、矿物质及多酚类化合物。所有参与者均在移动检查中心完成了两次 24 小时膳食回顾访谈, 两次访谈之间相隔 3 至 10 天。以两次访谈结果的平均值作为个体每日抗氧化剂摄入量的估计值。膳食数据的整理与量化工作借助美国农业部食物与营养研究数据库及其配套的黄酮数据库进行统一换算与整合。

2.3. CVD 和 CKD 定义

CVD 定义为参与者自报曾被医生诊断患有充血性心力衰竭、冠心病、心绞痛、心肌梗死或卒中中的任一种疾病。CKD 依据改善全球肾脏病预后组织(KDIGO)指南定义, 为肾脏结构或功能异常并对健康产生影响, 其诊断标准包括以下任一项: 估算肾小球滤过率(eGFR) $< 60 \text{ mL/min/1.73 m}^2$; 或尿白蛋白/肌酐比(UACR) $\geq 30 \text{ mg/g}$ [3]。其中, eGFR 采用 CKD-EPI 公式基于血清肌酐计算[12]。

2.4. 基线特征的收集

基线特征包括人口统计学因素、生活方式因素及健康状况。人口统计学因素包括年龄、性别、种族/族裔、教育程度、婚姻状况和家庭贫困收入比, 相关数据来源于 NHANES 人口统计学模块。

生活方式因素包括中高强度体力活动、饮酒情况、吸烟状态和膳食总能量摄入。中高强度体力活动根据工作活动及休闲活动问卷中高强度与中等强度活动的参与情况确定；饮酒情况依据各调查周期相应的饮酒问卷信息整理；吸烟状态根据终生是否吸烟至少 100 支及当前吸烟情况确定。膳食总能量摄入采用两次 24 小时膳食回顾中每日能量摄入量的平均值。

健康状况包括体质指数(body mass index, BMI)、糖尿病、高血压和高脂血症。BMI 来源于 NHANES 体格检查模块。糖尿病的判定综合考虑自报医生诊断、降糖药物使用、糖化血红蛋白、空腹血糖及口服葡萄糖耐量试验 2 小时血糖；高血压根据血压测量结果、自报诊断及降压药物使用情况确定；高脂血症根据甘油三酯、高密度脂蛋白胆固醇、自报高胆固醇诊断及降脂药物使用情况综合判定。相关信息来源于 NHANES 问卷、体格检查及实验室检测模块。

2.5. 机器学习预处理

建模前对候选变量的缺失情况进行检查，缺失比例超过 40%的变量不再纳入分析。针对 44 项膳食抗氧化剂，计算变量间的 Spearman 相关系数；当相关系数绝对值大于 0.90 时，删除其中冗余程度较高者。该项筛选仅涉及膳食抗氧化剂，不用于排除人口学、生活方式及健康状况变量。随后以 CKD 状态进行分层，按照 7:3 的比例划分训练集和测试集。连续变量采用中位数填补并进行 Z-score 标准化，分类变量采用众数填补后进行独热编码。预处理参数均在训练集及交叉验证过程中确定，再应用于测试集。主分析保留原始样本分布，不使用 SMOTE 或其他重采样方法。

2.6. 统计分析

在考虑 NHANES 复杂抽样设计的基础上，对单纯 CVD 组和 CVD 合并 CKD 组的描述性特征进行比较，复杂抽样设计因素包括抽样权重、分层和初级抽样单元。连续变量采用加权中位数和四分位数间距(interquartile range, IQR)进行描述，组间差异采用基于秩次的加权检验进行比较。分类变量采用未加权例数和加权百分比表示，组间差异采用 Rao - Scott 校正卡方检验进行评估。

在机器学习分析中，将三个调查周期的数据合并，并将调查周期作为协变量纳入模型。本研究构建逻辑回归(logistic regression, LR)、随机森林(random forest, RF)、极端随机树(extremely randomized trees, ET)、梯度提升树(gradient boosting machine, GBM)、XGBoost 和 LightGBM 六种分类模型。超参数调优在训练集中采用网格搜索结合分层 10 折交叉验证完成，并以 AUC-ROC 作为参数选择指标。对于支持样本权重的算法，在模型拟合过程中纳入 NHANES 调查权重，所有模型均在同一独立测试集中进行性能评价。

模型识别性能在独立测试集中进行评估，评价指标包括受试者工作特征曲线下面积(AUC-ROC)、精确率 - 召回率曲线下面积(AUC-PR)、Brier 评分、准确率、敏感度、特异度、精确率和 F1 评分。准确率、敏感度、特异度、精确率和 F1 评分均在默认分类阈值 0.50 下计算。F1 评分定义为精确率和敏感度的调和平均数；Brier 评分用于评价预测概率的准确性，其数值越低表示概率预测性能越好。进一步绘制 ROC 曲线、精确率 - 召回率曲线和校准曲线，以综合评价模型性能。所有机器学习模型均基于 scikit-learn 框架实现。

在完成模型性能评价后，对最优模型进一步进行可解释性分析。采用 SHapley Additive exPlanations (SHAP)方法解释模型输出，通过平均绝对 SHAP 值评估特征重要性，并利用 SHAP 摘要图及依赖图展示关键特征对模型输出的贡献方向及潜在非线性关系。

描述性统计分析采用 R 4.4.3 完成，复杂抽样分析基于 survey 包实现；机器学习建模及相关分析采用 Python 3.12 完成。所有统计检验均为双侧检验， $P < 0.05$ 认为差异有统计学意义。

3. 结果

3.1. 参与者基线特征

Table 1. Weighted baseline characteristics of CKD status grouping in CVD population
表 1. CVD 人群 CKD 状态分组的加权基线特征

变量	总体	单纯 CVD 组	CVD 合并 CKD 组	P 值
		加权 N = 7,497,010	加权 N = 3,888,752	
年龄(岁)	68.00 (57.00, 77.00)	63.00 (54.00, 71.00)	76.00 (68.00, 80.00)	<0.001
BMI (kg/m ²)	30.05 (26.06, 34.62)	29.90 (25.80, 34.00)	30.40 (26.09, 36.00)	0.497
总能量摄入(kcal/天)	1776.00 (1392.50, 2270.00)	1827.50 (1456.50, 2395.50)	1708.50 (1294.00, 2038.00)	<0.001
维生素 A	557.00 (357.50, 809.00)	544.00 (346.00, 791.00)	577.50 (372.00, 838.50)	0.591
维生素 C	62.80 (33.35, 116.20)	63.75 (31.25, 118.25)	59.05 (34.10, 114.15)	0.846
维生素 E	6.56 (4.65, 9.57)	6.75 (4.84, 9.74)	6.25 (4.35, 9.28)	0.092
镁	263.00 (195.50, 333.50)	272.50 (201.50, 353.00)	241.50 (184.50, 303.50)	<0.001
锌	10.05 (7.00, 13.61)	10.36 (7.22, 14.03)	9.60 (6.79, 13.13)	0.069
硒	96.60 (68.30, 128.60)	101.90 (69.40, 135.20)	88.65 (64.75, 113.75)	0.002
类胡萝卜素	6074.00 (2874.00, 12190.50)	6166.00 (3110.50, 12187.00)	5855.50 (2597.00, 12574.50)	0.493
儿茶素小计	12.87 (4.96, 65.40)	12.74 (4.76, 66.61)	13.19 (5.17, 62.55)	0.793
总异黄酮	0.01 (0.00, 0.07)	0.01 (0.00, 0.11)	0.01 (0.00, 0.04)	0.044
总花青素	2.28 (0.23, 17.43)	2.14 (0.15, 16.21)	2.89 (0.34, 18.08)	0.506
总黄烷-3-醇	13.64 (4.98, 158.76)	12.97 (4.76, 156.61)	14.15 (5.39, 160.19)	0.97
总黄烷酮	1.16 (0.05, 23.55)	1.27 (0.05, 23.78)	0.89 (0.05, 23.04)	0.762
总黄酮	0.54 (0.20, 1.07)	0.57 (0.23, 1.18)	0.46 (0.17, 0.91)	0.085
总黄酮醇	12.53 (7.30, 21.45)	12.25 (7.29, 22.08)	13.06 (7.31, 20.60)	0.636
29 种黄酮总量	65.17 (24.80, 223.48)	61.69 (24.80, 231.66)	69.85 (24.96, 220.93)	0.871
性别				0.161
男性	561 (57.8%)	335 (59.9%)	226 (53.6%)	
女性	385 (42.2%)	225 (40.1%)	160 (46.4%)	
种族/族裔				0.006
墨西哥裔美国人	88 (4.8%)	59 (5.0%)	29 (4.4%)	
其他西班牙裔	55 (3.7%)	37 (4.0%)	18 (3.1%)	
非西班牙裔白人	588 (77.0%)	354 (77.2%)	234 (76.7%)	
非西班牙裔黑人	175 (9.8%)	80 (7.7%)	95 (13.9%)	
其他种族	40 (4.7%)	30 (6.1%)	10 (1.9%)	
教育水平				0.442
高中以下	279 (21.1%)	159 (19.3%)	120 (24.8%)	
高中毕业	254 (28.5%)	156 (29.6%)	98 (26.1%)	
部分大学或副学士	259 (27.2%)	155 (27.3%)	104 (27.0%)	
大学及以上	154 (23.3%)	90 (23.8%)	64 (22.1%)	

续表

婚姻状态				0.035
已婚/与伴侣同居	546 (59.8%)	344 (63.7%)	202 (52.3%)	
分居	32 (3.3%)	23 (3.8%)	9 (2.3%)	
未婚/其他	368 (36.8%)	193 (32.4%)	175 (45.4%)	
家庭贫困收入比				0.086
≤1	185 (14.2%)	117 (15.2%)	68 (12.1%)	
1~3	473 (43.7%)	262 (40.3%)	211 (50.4%)	
>3	288 (42.1%)	181 (44.5%)	107 (37.5%)	
BMI 分层				0.531
体重过低	11 (1.0%)	6 (0.6%)	5 (1.8%)	
正常体重	177 (18.2%)	97 (18.4%)	80 (17.8%)	
超重	318 (30.2%)	194 (31.2%)	124 (28.3%)	
肥胖	440 (50.5%)	263 (49.7%)	177 (52.1%)	
吸烟状态				0.02
从不吸烟	367 (40.2%)	209 (36.8%)	158 (46.6%)	
既往吸烟	404 (43.5%)	233 (44.0%)	171 (42.6%)	
当前吸烟	175 (16.3%)	118 (19.2%)	57 (10.8%)	
中高强度体力活动				<0.001
否	452 (40.9%)	228 (33.9%)	224 (54.3%)	
是	494 (59.1%)	332 (66.1%)	162 (45.7%)	
饮酒频率				0.017
不饮酒	233 (21.2%)	122 (18.7%)	111 (25.9%)	
每月 1 至<5 次	568 (58.5%)	343 (56.8%)	225 (61.7%)	
每月 5 至<10 次	55 (8.9%)	33 (10.8%)	22 (5.3%)	
每月 ≥10 次	90 (11.4%)	62 (13.7%)	28 (7.1%)	
高血压				<0.001
否	185 (23.1%)	133 (28.8%)	52 (12.1%)	
是	761 (76.9%)	427 (71.2%)	334 (87.9%)	
高脂血症				0.505
否	153 (13.3%)	90 (12.7%)	63 (14.4%)	
是	793 (86.7%)	470 (87.3%)	323 (85.6%)	
糖尿病				0.001
否	561 (61.5%)	375 (68.0%)	186 (48.8%)	
是	385 (38.5%)	185 (32.0%)	200 (51.2%)	
调查周期				0.79
2007~2008	367 (35.4%)	214 (36.3%)	153 (33.5%)	
2009~2010	321 (32.2%)	192 (31.5%)	129 (33.4%)	
2017~2018	258 (32.5%)	154 (32.1%)	104 (33.1%)	

注：连续变量以中位数(Q1, Q3)表示，分类变量以未加权人数(%)表示。连续变量组间比较采用基于复杂抽样设计的Kruskal-Wallis 检验，分类变量组间比较采用 Rao-Scott 校正 Pearson 卡方检验。

本研究共纳入 946 名 CVD 参与者, 其中单纯 CVD 组 560 人, CVD 合并 CKD 组 386 人。膳食指标方面, CVD 合并 CKD 组的总能量、镁和硒摄入量低于单纯 CVD 组(均 $P < 0.05$), 两组总异黄酮摄入分布也存在差异($P = 0.044$), 其余维生素、矿物质、类胡萝卜素及黄酮类指标在两组间未见明显差别。除膳食特征外, CVD 合并 CKD 组年龄较大, 中高强度体力活动比例较低, 高血压和糖尿病比例较高。两组在种族/族裔、婚姻状态、吸烟状态和饮酒频率方面亦存在差异, 而性别、教育水平、家庭贫困收入比、BMI、高血脂症及调查周期分布较为接近(表 1)。

3.2. 膳食抗氧化剂相关性筛选



Figure 2. Correlation matrix of dietary antioxidants retained after excluding collinearity
图 2. 排除共线性后保留的膳食抗氧化剂的相关矩阵

在机器学习模型构建前, 对 44 项膳食抗氧化剂摄入量进行了 Spearman 相关分析, 相关矩阵见附录图 S1。结果显示, 部分膳食抗氧化剂之间相关程度较高。经相关性筛选, 剔除矮牵牛素、茶黄素、茶红素、茶黄素-3,3'-双没食子酸酯、茶黄素-3'-没食子酸酯、儿茶素小计、总异黄酮、总黄烷酮、总黄酮和总黄酮醇,

共 10 项变量。最终保留 34 项膳食抗氧化剂用于后续机器学习建模及 SHAP 分析，筛选后的相关矩阵。最终保留 34 项膳食抗氧化剂用于后续机器学习建模及 SHAP 分析，筛选后的 Spearman 相关矩阵见图 2。

3.3. 机器学习模型性能

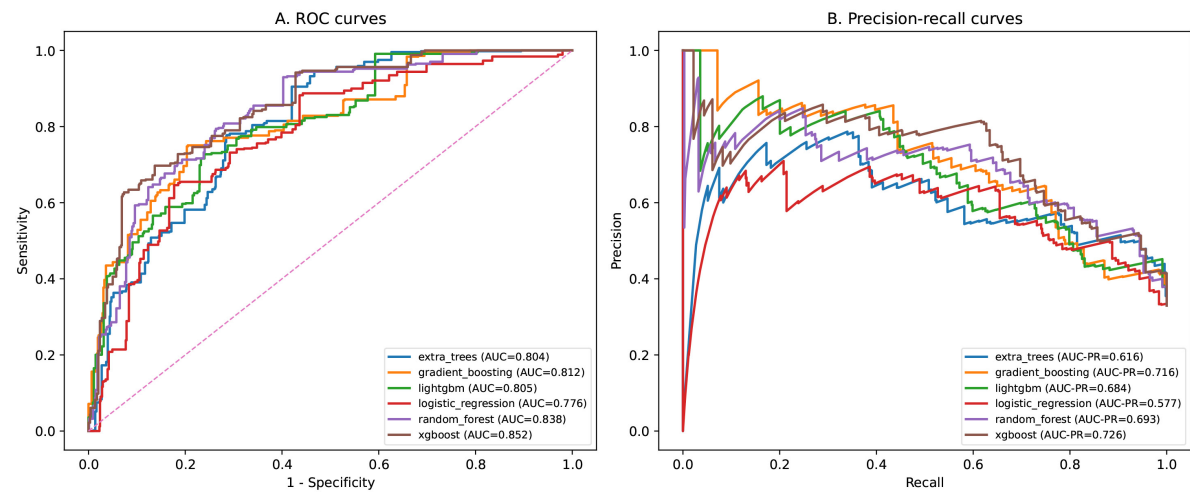
表 2 总结了 6 种机器学习模型在 CVD 人群中识别 CKD 的加权性能，各模型在测试集上的 AUC-ROC 介于 0.776~0.852 之间。XGBoost 的 AUC-ROC、AUC-PR 和敏感度分别为 0.852、0.726 和 0.781，均为六种模型中最高。随机森林的准确率和 F1 评分最高，分别为 0.787 和 0.677，整体表现较为均衡。梯度提升树的特异度和精确率最高，Brier 评分最低，但敏感度相对较低。综合多项指标，XGBoost 表现相对较好，因此用于后续 SHAP 分析。

Table 2. Performance of different machine learning models in identifying CKD in CVD populations
表 2. 不同机器学习模型在 CVD 人群中识别 CKD 的性能

模型	准确率	敏感度	特异度	精确率	F1 评分	AUC-ROC	AUC-PR	Brier 评分
XGBoost	0.742	0.781	0.723	0.581	0.666	0.852	0.726	0.184
随机森林	0.787	0.675	0.842	0.678	0.677	0.838	0.693	0.167
梯度提升树	0.781	0.555	0.892	0.717	0.626	0.812	0.716	0.159
LightGBM	0.736	0.593	0.806	0.601	0.597	0.805	0.684	0.174
极端随机树	0.709	0.673	0.727	0.549	0.605	0.804	0.616	0.179
逻辑回归	0.716	0.686	0.731	0.557	0.615	0.776	0.577	0.188

注：分类阈值均为 0.50；F-beta 中的 $\beta = 1$ ，与 F1 分数等价。所有数值均为采用 NHANES 调查权重计算的测试集结果。

各模型 ROC 曲线(图 3(A))与 PR 曲线(图 3(B))的趋势与数值结果一致。ROC 曲线中，XGBoost 和随机森林的曲线总体位置相对靠上，梯度提升树、LightGBM 和极端随机树的曲线较为接近，逻辑回归相对较低；PR 曲线中，XGBoost 和梯度提升树表现相对较好，随机森林和 LightGBM 次之，极端随机树和逻辑回归相对较低。综合多项评价指标，XGBoost 在整体区分能力和阳性个体识别方面表现相对较好，因此选取该模型进一步开展 SHAP 可解释性分析。

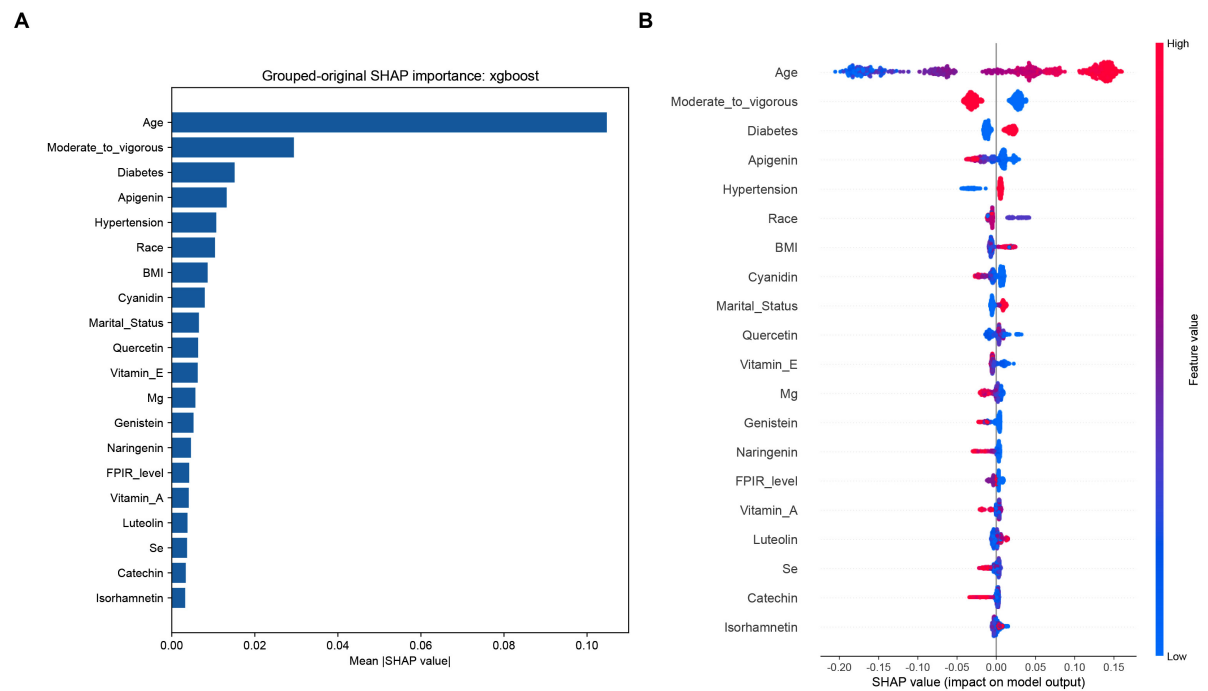


注：A 为 ROC 曲线；B 为 PR 曲线。

Figure 3. ROC curves and PR curves of different machine learning models on the test set
图 3. 不同机器学习模型在测试集中的 ROC 曲线和 PR 曲线

3.4. SHAP 解释与特征重要性分析

XGBoost 模型的 SHAP 分析结果见图 4。平均绝对 SHAP 值排序显示，年龄对模型输出的贡献最高，其后为中高强度体力活动、糖尿病、芹菜素、高血压、种族/族裔和 BMI。膳食抗氧化剂中，芹菜素的贡献最高；其余进入前 20 位的膳食抗氧化剂包括矢车菊素、槲皮素、维生素 E、镁、染料木素、柚皮素、维生素 A、木犀草素、硒、儿茶素和异鼠李素。SHAP 摘要图显示，高龄、糖尿病和高血压多对应正 SHAP 值，缺乏中高强度体力活动也主要分布于正 SHAP 值一侧；较高的芹菜素和矢车菊素摄入则多对应负 SHAP 值。正 SHAP 值表示相关特征推动模型将参与者识别为 CVD 合并 CKD，负 SHAP 值表示模型更倾向于将其识别为单纯 CVD。



注：A，基于平均绝对 SHAP 值的特征重要性排序；B，SHAP 摘要图。

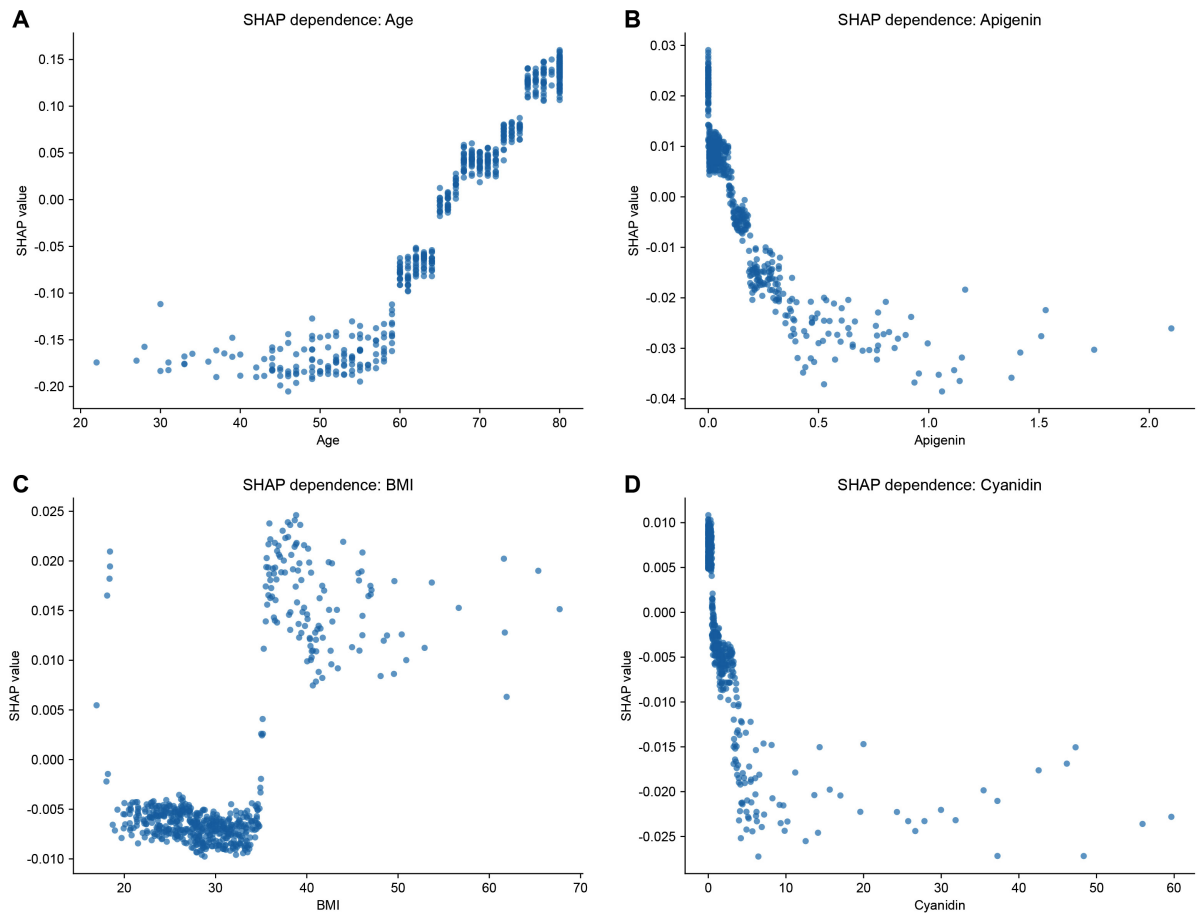
Figure 4. Explanation of key features of the SHAP-based XGBoost model
图 4. 基于 SHAP 的 XGBoost 模型关键特征解释

SHAP 依赖图见图 5。年龄的 SHAP 值在约 60 岁以后明显上升，并在 65 岁左右由负值转为正值，此后随年龄增加继续升高。芹菜素的 SHAP 值在低摄入区间由正值迅速转为负值，摄入量进一步增加后总体维持在负值区间，变化逐渐趋缓。矢车菊素的变化趋势与之类似，其 SHAP 值随摄入量增加迅速下降，随后在负值区间趋于平稳。BMI 与模型输出之间呈明显的非线性关系，约 18~34 kg/m² 范围内多数样本对应负 SHAP 值，35 kg/m² 以上则主要表现为正 SHAP 值，少数较低 BMI 参与者也呈现正 SHAP 值。

4. 讨论

本研究基于 NHANES 三个调查周期的数据，采用可解释机器学习方法识别 CVD 人群中合并 CKD 的个体。结果显示，XGBoost 在六种模型中的 AUC-ROC、AUC-PR 和敏感度最高，分别为 0.852、0.726 和 0.781，在整体区分能力和阳性个体识别方面表现较好。SHAP 分析显示，年龄、中高强度体力活动、糖尿病、高血压和 BMI 等基础特征对模型输出贡献较高；在膳食抗氧化剂中，芹菜素、矢车菊素、槲皮

素、维生素 E 和镁等的 SHAP 重要性相对较高。年龄、BMI、芹菜素和矢车菊素与模型输出之间还呈现出不同程度的非线性关系。上述结果表明, CVD 人群中的 CKD 状态差异可能同时涉及年龄、生活方式、代谢性疾病和膳食摄入等多方面信息。



注: A, 年龄; B, 芹菜素; C, BMI; D, 矢车菊素。图中 SHAP 值表示对应变量对模型输出结果的贡献方向及大小, 颜色表示交互变量取值。

Figure 5. SHAP dependence graph of age, apigenin, BMI, and cyanidin
图 5. 年龄、芹菜素、BMI、矢车菊素的 SHAP 依赖图

既往研究多从一般成年人或特定代谢性疾病人群出发, 分析膳食抗氧化剂与 CVD、CKD 之间的关系, 而专门针对 CVD 人群中 CKD 状态差异的研究相对较少。年龄是本研究中贡献最高的特征, 高血压和糖尿病也多推动模型将参与者识别为 CVD 合并 CKD, 这与心血管、肾脏和代谢异常之间相互影响的既有认识一致[1]-[3]。缺乏中高强度体力活动同样具有较高贡献, 但该变量也可能同时反映活动能力下降、基础疾病较重或总体健康状况较差。氧化应激和慢性炎症是 CVD 与 CKD 共同涉及的病理过程, 为膳食抗氧化成分参与模型识别提供了一定生物学依据[4]。既往 NHANES 研究发现, 较高的复合膳食抗氧化指数分别与较低的 CKD 和 CVD 患病可能性相关[5][6]; 但在一项 CKD 队列研究中, 氧化平衡评分与肾脏及心血管结局的关联在充分调整后有所减弱[7]。这些并不完全一致的结果也说明, 膳食抗氧化指标的意义可能受到研究人群、结局类型和其他健康特征的影响。

本研究中, CVD 合并 CKD 组的镁和硒摄入量低于单纯 CVD 组, 其中镁也进入了 SHAP 重要特征序

列。前瞻性研究曾报告较低膳食镁摄入与较高的新发 CKD 风险相关[13]。不过, CVD 合并 CKD 组的总能量摄入同样较低, 因此镁和硒的组间差异也可能与总体食物摄入、膳食结构、年龄和疾病状态有关。芹菜素是模型中贡献最高的膳食抗氧化剂, 其较高取值多对应负 SHAP 值。动物和细胞研究显示, 芹菜素可能参与氧化应激、炎症及细胞凋亡等过程的调节[14]; 关于糖尿病肾病的非临床研究综述也观察到类似作用, 但临床证据仍然有限[15]。矢车菊素在本研究中的变化趋势与芹菜素相近, 既往实验研究发现, 矢车菊素-3-葡萄糖苷可能影响糖尿病肾损伤中的氧化应激和炎症通路[16], 并可能参与肾纤维化相关过程[17]。这些实验结果只能为本研究的模型模式提供机制上的解释, 不能证明增加芹菜素或矢车菊素摄入可以预防 CKD。SHAP 负值仅表示相应取值使模型更倾向于识别为单纯 CVD, 并不等同于保护效应。

从模型性能来看, 选择 XGBoost 作为主要解释模型, 主要依据其 AUC-ROC、AUC-PR 和敏感度相对较高, 并不意味着该模型在所有指标上均优于其他算法。随机森林的准确率和 F1 评分更高, 梯度提升树则具有更高的特异度和精确率, 以及更低的 Brier 评分, 表明不同模型在性能维度上各有侧重。此前两项膳食抗氧化剂机器学习研究得到的最优模型也存在差异, CVD 与癌症共病研究中 LightGBM 表现较好, 而糖尿病人群 CVD 识别研究中 XGBoost 表现较好[8] [9]。已有 CKD 研究也表明, 不同机器学习算法在疾病识别或进展评估中的性能会受到样本结构、结局定义和候选特征的影响[10] [11]。在临床应用方面, 本研究中的膳食抗氧化剂摄入量需基于两次 24 小时膳食回顾并结合食物成分数据库计算, 在常规临床实践中较难快速获取, 且可能受到回忆偏倚、日间摄入波动及地区食物构成差异的影响, 从而限制模型的直接应用。因此, 该模型现阶段不宜直接作为临床诊断或筛查工具。未来可在外部验证的基础上筛选稳定且易于获取的膳食特征, 开发简化的膳食抗氧化剂问卷或电子化评估工具, 并评价其在常规临床特征之外的补充识别价值。

本研究基于具有复杂抽样设计的 NHANES 数据, 综合纳入人口学、生活方式、健康状况及膳食抗氧化剂特征, 并通过多模型比较和 SHAP 分析提高了结果的可解释性。但本研究仍存在一定局限。首先, 横断面设计无法确定相关特征与 CKD 状态之间的时间顺序和因果关系, 本研究模型仅用于识别参与者在调查时是否合并 CKD, 不能用于预测未来 CKD 的发生或进展。其次, CKD 依据单次 eGFR 和 UACR 测量判定, CVD 及部分生活方式资料来源于自报, 可能存在一定误分类。再次, 两次 24 小时膳食回顾难以完全反映长期膳食摄入, 并可能受到回忆偏倚和日间摄入波动的影响。此外, 研究样本量有限, 模型尚缺少独立外部验证。尽管本研究已通过 Spearman 相关分析剔除高度相关的膳食抗氧化剂, 保留特征之间仍可能存在一定相关性。相关特征的 SHAP 贡献可能相互分配或替代, 从而影响单个抗氧化剂 SHAP 值及重要性排序的稳定性。因此, 本文的 SHAP 排序仅反映各特征在当前模型中的相对重要性, 对单个抗氧化剂的重要性排名应谨慎解释。

综上, XGBoost 能够较好地地区分单纯 CVD 与 CVD 合并 CKD 参与者, 部分膳食抗氧化剂可在年龄、体力活动及代谢性疾病等基础特征之外提供补充识别信息。未来需在外部样本中进一步验证模型的稳定性, 并通过前瞻性队列研究评估膳食抗氧化剂与 CKD 发生、进展及肾功能变化之间的纵向关联。

声 明

NHANES 研究方案已获得国家健康统计中心伦理审查委员会批准, 所有参与者均签署书面知情同意。本研究仅使用公开可获取的数据, 因此无需额外伦理审批。

参考文献

- [1] Ndumele, C.E., Rangaswami, J., Chow, S.L., Neeland, I.J., Tuttle, K.R., Khan, S.S., *et al.* (2023) Cardiovascular-Kidney-Metabolic Health: A Presidential Advisory from the American Heart Association. *Circulation*, **148**, 1606-1635.

- <https://doi.org/10.1161/cir.0000000000001184>
- [2] Rangaswami, J., Bhalla, V., Blair, J.E.A., Chang, T.I., Costa, S., Lentine, K.L., *et al.* (2019) Cardiorenal Syndrome: Classification, Pathophysiology, Diagnosis, and Treatment Strategies: A Scientific Statement from the American Heart Association. *Circulation*, **139**, e840-e878. <https://doi.org/10.1161/cir.0000000000000664>
 - [3] Kidney Disease: Improving Global Outcomes CKD Work Group (2024) KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. *Kidney International*, **105**, S117-S314.
 - [4] Ravarotto, V., Simioni, F., Pagnin, E., Davis, P.A. and Calò, L.A. (2018) Oxidative Stress—Chronic Kidney Disease—Cardiovascular Disease: A Vicious Circle. *Life Sciences*, **210**, 125-131. <https://doi.org/10.1016/j.lfs.2018.08.067>
 - [5] Wang, M., Huang, Z.H., Zhu, Y.H., He, P. and Fan, Q.L. (2023) Association between the Composite Dietary Antioxidant Index and Chronic Kidney Disease: Evidence from NHANES 2011-2018. *Food & Function*, **14**, 9279-9286. <https://doi.org/10.1039/d3fo01157g>
 - [6] Wang, R., Tao, W. and Cheng, X. (2024) Association of Composite Dietary Antioxidant Index with Cardiovascular Disease in Adults: Results from 2011 to 2020 NHANES. *Frontiers in Cardiovascular Medicine*, **11**, Article 1379871. <https://doi.org/10.3389/fcvm.2024.1379871>
 - [7] Ilori, T.O., Wang, X., Huang, M., Gutierrez, O.M., Narayan, K.M.V., Goodman, M., *et al.* (2017) Oxidative Balance Score and the Risk of End-Stage Renal Disease and Cardiovascular Disease. *American Journal of Nephrology*, **45**, 338-345. <https://doi.org/10.1159/000464257>
 - [8] Qi, X., Wang, S., Fang, C., Jia, J., Lin, L. and Yuan, T. (2025) Machine Learning and SHAP Value Interpretation for Predicting Comorbidity of Cardiovascular Disease and Cancer with Dietary Antioxidants. *Redox Biology*, **79**, Article 103470. <https://doi.org/10.1016/j.redox.2024.103470>
 - [9] Zhang, X., Lin, S., Zeng, Q., Peng, L. and Yan, C. (2025) Machine Learning and SHAP Value Interpretation for Predicting Cardiovascular Disease Risk in Patients with Diabetes Using Dietary Antioxidants. *Frontiers in Nutrition*, **12**, Article 1612369. <https://doi.org/10.3389/fnut.2025.1612369>
 - [10] Ghosh, S.K. and Khandoker, A.H. (2024) Investigation on Explainable Machine Learning Models to Predict Chronic Kidney Diseases. *Scientific Reports*, **14**, Article No. 3687.
 - [11] Takkavatakarn, K., Oh, W., Cheng, E., *et al.* (2023) Machine Learning Models to Predict End-Stage Kidney Disease in Chronic Kidney Disease Stage 4. *BMC Nephrology*, **24**, Article No. 376.
 - [12] Inker, L.A., Eneanya, N.D., Coresh, J., Tighiouart, H., Wang, D., Sang, Y., *et al.* (2021) New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race. *New England Journal of Medicine*, **385**, 1737-1749. <https://doi.org/10.1056/nejmoa2102953>
 - [13] Koh, H.B., Kim, H.J., Heo, G.Y., Kim, H.W., Jung, C., Han, S.H., *et al.* (2024) Association between Dietary Magnesium Intake and Incident Chronic Kidney Disease: A Prospective Observational Cohort Study. *The American Journal of Clinical Nutrition*, **120**, 964-972. <https://doi.org/10.1016/j.ajcnut.2024.08.009>
 - [14] Wu, Q., Li, W., Zhao, J., Sun, W., Yang, Q., Chen, C., *et al.* (2021) Apigenin Ameliorates Doxorubicin-Induced Renal Injury via Inhibition of Oxidative Stress and Inflammation. *Biomedicine & Pharmacotherapy*, **137**, Article 111308. <https://doi.org/10.1016/j.biopha.2021.111308>
 - [15] Muhammed, T.M., Jalil, A.T., Taher, W.M., *et al.* (2024) The Effects of Apigenin in the Treatment of Diabetic Nephropathy: A Systematic Review of Non-Clinical Studies. *Mini-Reviews in Medicinal Chemistry*, **24**, 341-354. <https://doi.org/10.2174/1389557523666230811092423>
 - [16] Zheng, H.X., Qi, S.S., He, J., Hu, C.Y., Han, H., Jiang, H., *et al.* (2020) Cyanidin-3-Glucoside from Black Rice Ameliorates Diabetic Nephropathy via Reducing Blood Glucose, Suppressing Oxidative Stress and Inflammation, and Regulating Transforming Growth Factor β 1/Smad Expression. *Journal of Agricultural and Food Chemistry*, **68**, 4399-4410. <https://doi.org/10.1021/acs.jafc.0c00680>
 - [17] Lu, Q., Liu, J., Xiong, Y., Jian, J., Wang, J., Chen, Z., *et al.* (2025) Cyanidin-3-Glucoside Upregulated NDRG2 through the PI3K/AKT Pathway to Alleviate EMT and ECM in Renal Fibrosis. *Scientific Reports*, **15**, Article No. 10695. <https://doi.org/10.1038/s41598-025-94918-x>

144

