·创刊40周年·

代谢组流行病学研究进展

杭栋　沈洪兵
南京医科大学公共卫生学院流行病学系 211166
通信作者：沈洪兵，Email：hbshen@njmu.edu.cn

【摘要】 近年来，基于代谢组学技术平台和数据分析方法的快速发展，作为系统流行病学的重要分支——代谢组流行病学正获得越来越多的关注。代谢组流行病学有助于更好地描述暴露特征，反映环境-基因相互作用的效应，阐明暴露与疾病的“黑箱”机制，并发现新的生物标志物。本文简要介绍代谢组流行病学研究的定义、方法、研究进展及展望。

【关键词】 系统流行病学；　代谢组流行病学；　糖尿病；　心血管疾病；　肿瘤
基金项目：国家自然科学基金（81521004，81973127）；江苏省自然科学基金（BK20190083）

Progress in metabolomics epidemiology

Hang Dong, Shen Hongbing
Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing 211166, China
Corresponding author: Shen Hongbing, Email: hbshen@njmu.edu.cn

【**Abstract**】 In recent years, with the rapid development of metabonomic technology and analytic methods, metabolomics epidemiology, as an important branch of systems epidemiology, attracts more attention. Metabolomics epidemiology can better describe the characteristics of exposure, reflect the interaction between environmental factors and genetics, uncover the "black box" of the mechanisms underlying exposure and disease, and identify new biomarkers. This article briefly introduces the definition, methods, and progress of metabolomics epidemiology.

【**Key words**】 Systems epidemiology; Metabolomics epidemiology; Diabetes; Cardiovascular disease; Cancer
Fund programs: National Natural Science Foundation of China (81521004, 81973127); Natural Science Foundation of Jiangsu Province (BK20190083)

系统流行病学强调传统流行病学与现代高通量多组学技术的有机整合，是现代病因学研究和精准预防的重要方向[1]。作为基因组、转录组和蛋白质组之后兴起的新的组学研究热点，代谢组学主要采用核磁共振（nuclear magnetic resonance，NMR）或质谱（mass spectrometry，MS）技术，对生物体内参与生化反应的数百至数千种中间产物及终产物（如氨基酸、脂类、有机酸等）进行定性和定量分析，描述内源性代谢物质的整体特征及其对内因和外因变化的应答规律，在疾病诊断、药物研发、生物学功能研究等领域应用广泛[2]。代谢组流行病学整合了代谢组学技术与流行病学研究方法，是系统流行病学的重要组成部分，相关研究的数量正逐步增加[3]。本文简要介绍代谢组流行病学的定义和内涵、研究设计、检测方法以及常用的数据处理和分析方法，同时也将介绍代谢组学在糖尿病、心血管疾病、恶性肿瘤流行病学研究中的进展及展望。

一、代谢组流行病学的定义和内涵

迄今为止，国内外文献尚未见代谢组流行病学的定义和内涵。笔者建议给出如下定义：代谢组流

DOI：10.3760/cma.j.cn112338-20210413-00310
收稿日期　2021-04-13　　**本文编辑**　李银鸽
引用本文：杭栋，沈洪兵．代谢组流行病学研究进展[J]．中华流行病学杂志，2021，42(7)：1148-1153．DOI：10.3760/cma.j.cn112338-20210413-00310.

行病学是应用流行病学与代谢组学相结合的方法，系统研究人群中代谢物的分布及其与疾病/健康的关系和影响因素，评价代谢物在防治疾病、促进健康方面的应用价值。

代谢组流行病学是系统流行病学的重要组成部分。与其他组学相比，代谢组处于生命网络调控的下游，可反映环境和基因交互作用的末端效应，更接近于反映人体生理或病理状况；基因组和蛋白质组在功能水平上的微小变化可通过代谢过程得到放大，而非功能性变化则一般不会在代谢产物上得到反映，因此，代谢组流行病学研究有助于发现影响健康状况的关键事件，对于探索疾病发生发展的规律具有不可替代的作用，是实现系统流行病学——深入理解多层次多因素间复杂的调控网络及其相互作用，进行“暴露因素-组学标志物-疾病结局”病因学推断，建立疾病发生风险预警模型的关键[4]。

二、代谢组流行病学的研究设计、检测技术和统计分析方法

1. 研究设计：根据研究的目的不同，可采用不同的流行病学研究设计，分为观察性（如队列研究、巢式病例对照研究、传统病例对照研究）和干预实验（如随机对照试验）。

队列研究设计是在基线时测量所有研究对象的代谢产物，追踪观察并比较不同代谢物水平组中结局发生率的差异，从而判定代谢物与结局有无关联及关联强度。由于队列研究的前瞻性特点符合因果推断的必要前提——时间顺序，在排除偏倚和混杂后可提供较高等级的人群证据。代谢组流行病学的一个发展趋势就是在已建立的、长期随访的高质量人群队列中开展相关代谢组学研究，推动病因学机制研究、生物标志物的发现及潜在干预靶点的识别。

巢式病例对照研究是在队列内套用病例对照研究的一种设计，以队列中所有的目标病例作为病例组，再根据病例发病时间，从同一队列的未发病者中随机匹配一个或多个对照，组成对照组。由于病例和对照的暴露（代谢组）在结局发生前获得，故一般不存在暴露与结局的时间顺序问题，且病例和对照来自同一队列，可比性较好。涉及的样本量小于队列研究，节约人力物力，因此在代谢组流行病学中较为常用。

传统病例对照研究设计是检验暴露与疾病相关性的快速方法，易于组织实施，但对照的选择比较复杂，容易发生选择偏倚。此外，病例组生物标本采集时已经发生疾病，代谢物水平可能受疾病状态的影响，导致反向因果关系，因此该设计的因果论证强度受限。但通过比较疾病早期患者与健康参与者的代谢物水平，有可能发现差异的代谢物，为提高早期诊断水平提供新的生物标志物，具有二级预防意义。

随机对照试验是将研究对象随机分配到干预组和对照组，人为施加或减少某种处理因素后，随访观察处理因素的作用效果。结合代谢组学技术，目前主要用于评估如饮食、运动、减重等干预措施对体内代谢物水平的影响，从而鉴定饮食特定标志物或反映干预效果的客观指标。与观察性研究相比，高质量的随机对照试验不容易出现混杂，因果论证强度更高，但也容易受到干预依从性差、持续时间有限、人群代表性不足、费用昂贵等问题的影响。

2. 检测技术：代谢组学的快速发展得益于近十几年来仪器联用技术和数据挖掘技术的迅速发展。目前常用的有 3 个技术平台：NMR、气相色谱-质谱（gas chromatography-mass spectrometry，GC-MS）和液相色谱-质谱（liquid chromatography-mass spectrometry，LC-MS）。一般而言，NMR 无需提取等预处理步骤即可检测样本中存在的代谢物，但与 MS 方法相比灵敏度较差，启动成本高；GC-MS 灵敏度较高，成本较低，但局限于检测小分子挥发性物质，不适用于检测热不稳定或难挥发的化合物；LC-MS 灵敏度高，检测物质分子量范围广，尤其是高沸点、大分子、强极性或热稳定性差的化合物[5-7]。

代谢组学检测有靶向和非靶向两种方法。靶向方法采用内标化合物定量一组预先确定的代谢物，该方法具有较高的特异性和准确性，被广泛应用于不同生理状态下特定代谢产物的分析和比较。非靶向方法理论上是对样品中所有可测代谢物的综合检测，包括未知代谢物，因此在广泛识别新的代谢途径和生物标志物方面具有强大的潜力。早期流行病学研究大多采用靶向方法，近年来在大规模队列研究中应用非靶向方法的情况有所增加。如美国的护士健康研究（Nurses′ Health Study Ⅰ/Ⅱ）、医疗专业人员随访研究（Health Professionals Follow-Up Study）、妇女健康倡议研究（Women′s Health Initiatives）及前列腺、肺、结直肠和卵巢癌症筛查试验（Prostate，Lung，Colorectal and Ovarian

Cancer Screening Trial, PLCO)；西班牙的地中海饮食预防研究(Prevención con Dieta Mediterránea, PDM)；英国双胞胎队列(TwinsUK)等均采用了非靶向LC-MS进行代谢组检测。此外，美国弗雷明汉心脏研究(Framingham Heart Study, FHS)采用了高通量靶向LC-MS，中国慢性病前瞻性研究(China Kadoorie Biobank, CKB)和英国生物银行(UK Biobank)采用了高通量靶向NMR技术。

3. 统计学方法：在应用统计方法之前，代谢组学原始检测数据要经过预处理。MS数据预处理包括基线校正、保留时间对齐、谱峰检测与识别、积分等操作，而NMR数据预处理还包括谱图去噪、相位校正和定标等操作[8-9]。之后图谱信息被转换成统一格式的数据集，经过归一化(normalization)、尺度化(scaling)及数据转换(transformation)等处理后才能进行后续分析[8]。归一化的目的是消除检测过程中任何不必要的误差(如实验批次效应)；尺度化是通过调整数据的方差结构，改善后续的多变量统计分析的结果；数据转换是将偏态分布的代谢组学数据转换成正态分布，以满足线性分析的要求。

需要注意的是，代谢组学数据集通常含有缺失值。这有可能是由于生物学因素，如药物代谢产物在未服药者中缺失，也可能是检测技术的限制，如低强度信号无法与背景分离、信号强度低于仪器检测下限、仪器性能不稳定造成的检测误差等[10]。目前常用的处理方法包括用零、最小检测值的一半(或特定的比例)进行缺失填补，或采用复杂的统计方法，如k最近邻(k-nearest neighbors)、贝叶斯模型(Bayesian model)、主成分分析(principal component analysis, PCA)、随机森林(random forest)、基于链式方程的多重插补(multiple imputation by chained equations, MICE)等[11-13]。

代谢组流行病学的统计方法可分为单变量和多变量分析。前者主要用于快速考察各个代谢物在不同组别之间的差异，如t检验和秩和检验，也可应用logistic回归或Cox回归计算比值比(odds ratio, *OR*)或相对危险度(relative risk, *RR*)，反映代谢物与结局的关联强度。在对成百上千种代谢物进行单变量分析时，需要校正多重假设检验以降低Ⅰ类错误的发生率。传统的校正方法有Bonferroni和FDR(false discovery rate)，但由于代谢物之间存在较高的相关性，这些方法通常过于保守。替代的方法有置换检验(permutation test)，通过估计零假设下的P值分布，从而得到与样本类型和检测方法相适应的阈值[14]。由于代谢组学产生的是高维数据，需要多变量分析方法以揭示变量间复杂的相互关系，常用的方法包括PCA、偏最小二乘判别分析(partial least squares discrimination analysis, PLS-DA)、正交偏最小二乘法判别分析(Orthogonal PLS-DA)、聚类分析、通路分析、富集分析、随机森林，以及新方法如LASSO回归和网络分析等[9]。此外，可利用代谢物构建预测模型，通过C-statistics、NRI(net reclassification improvement)、IDI(integrated discrimination improvement)等方法评价模型的优劣[15]。

三、代谢组流行病学的应用

近年来国内外学者基于人群队列开展了多项代谢组流行病学研究，涉及糖尿病、心脑血管疾病、恶性肿瘤等常见慢性疾病。本文选择部分代表性研究进行介绍。

1. 糖尿病代谢组流行病学：目前已有不少研究报道了糖尿病代谢组学研究成果[16]。Merino等[17]在美国弗雷明汉心脏研究后代队列的1 150名参与者中，前瞻性分析了LC-MS靶向检测的220个血浆代谢物与糖尿病发生率的关联，发现甘氨酸和牛磺酸水平升高与糖尿病的发生风险降低有关，而苯丙氨酸增加糖尿病的风险；将19个代谢物加入传统因素模型中可显著提高糖尿病的风险预测能力。华中科技大学邬堂春教授研究团队基于同济东风队列和江苏非传染性疾病队列开展巢式病例对照研究，在1 559对年龄(±5岁)和性别匹配的糖尿病患者和对照中应用LC-MS靶向检测了52个血浆代谢物，发现丙氨酸、苯丙氨酸、酪氨酸和棕榈酰肉碱的水平升高与糖尿病发病风险增加有关，其中棕榈酰肉碱与糖尿病的关联为首次报道，为揭示糖尿病发生机制提供了新线索[18]。中国南京医科大学与美国哈佛大学研究团队合作，探讨了长期饮用咖啡的代谢谱与糖尿病的风险关联[19]：首先基于1 595名女性的非靶向LC-MS检测的血浆代谢组数据，鉴定出与咖啡的摄入量相关的34个代谢物，继而通过巢式病例对照研究发现15个咖啡相关代谢物与糖尿病风险有关，提示咖啡预防糖尿病的潜在机制；将相关代谢物加入传统危险因素预测模型，有助于提高糖尿病的风险预测水平，具有潜在应用价值[19]。

2. 心血管病代谢组流行病学：美国哈佛大学研究团队通过运用LC-MS非靶代谢组学技术对西班牙和美国的多个队列进行研究，采用弹性网络回归

(elastic net regression)方法从302个血浆代谢物中鉴定到67个代谢物与地中海饮食评分显著相关;较高评分相关的代谢物(如高不饱和脂质)与较低心血管病(CVD)风险有关,较低评分相关的代谢物(如谷氨酸)与更高CVD风险有关;孟德尔随机化分析支持上述代谢特征与心血管风险的关联[20]。因此,血浆代谢组可用于评估个体对地中海饮食的代谢反应,有助于预测未来患心血管疾病的风险。CKB项目团队采用NMR靶向测定了巢式病例对照研究中4 660名研究对象的225个代谢物,发现脂蛋白和脂质与心肌梗死和缺血性脑卒中的关联相似,但与出血性脑卒中无关;高密度脂蛋白颗粒与心肌梗死呈负相关,而TG与心肌梗死呈正相关;糖蛋白乙酰、酮体、葡萄糖和二十二碳六烯酸与上述疾病均相关,结果有助于深入研究心脑血管发病机制及鉴定相关生物标志物[21]。

3. 肿瘤代谢组流行病学:

(1)肺癌:Seow等[22]在中国上海女性健康研究(Shanghai Women's Health Study)队列中采用LC-MS非靶向检测了非吸烟女性275名肺癌病例和289名对照的尿液代谢组,发现高水平5-甲基-2-呋喃甲酸与肺癌发病风险降低相关;通路分析提示一碳代谢、核苷酸代谢、氧化应激和炎症可能参与非吸烟女性肺癌的发生。此外,Wen等[23]针对美国安德森癌症中心的386例肺癌患者和193例对照,采用非靶向和靶向相结合的方法发现血清胆红素水平在两组中存在显著差异,继而在42万余名参与者的前瞻性队列中进行验证,发现较低水平的胆红素与男性吸烟者肺癌发病率和死亡率风险升高有关。

(2)乳腺癌:在一项基于欧洲癌症前瞻性调查(European Prospective Investigation into Cancer, EPIC)的巢式病例对照研究中,研究者采用靶向技术检测了1 624对乳腺癌病例和相匹配对照的127个代谢物,分析发现在基线未使用雌激素的女性中,血浆高水平酰基肉碱C2与乳腺癌的发病风险增加有关,而磷脂酰胆碱、精氨酸和天冬酰胺与乳腺癌的风险降低有关[24]。来自美国PLCO的乳腺癌巢式病例对照研究(621对绝经后乳腺癌病例和相匹配对照)发现,617个血清代谢物中有67个与BMI密切相关,其中16a-羟基-脱氢异雄酮-3-硫酸盐和3-甲基戊二酰肉碱的水平升高增加乳腺癌发生风险;中介分析(mediation analysis)发现这两个代谢物介导了BMI与乳腺癌57.6%的关联效应,提示了肥胖促进乳腺癌发生的代谢途径[25]。近期的另一项巢式病例对照研究基于美国癌症预防研究(Cancer Prevention Study-Ⅱ)队列,采用非靶向LC-MS检测了782对绝经后乳腺癌病例和相匹配对照的1 275个血清代谢物,结果显示9个代谢物与乳腺癌呈正相关,15个代谢物与乳腺癌呈负相关,主要涉及肉碱、甘油酯和性类固醇代谢物,为阐明乳腺癌的代谢异常机制提供了更多线索[26]。

(3)结直肠癌:尽管已有较多研究探讨结直肠癌的代谢异常[27],但绝大多数采用非前瞻性的研究设计(横断面或传统病例对照研究)或临床组织标本(如癌和癌旁),在揭示代谢物与结直肠癌因果关联及转化应用方面存在局限性。近期来自EPIC队列的一项研究发现,世界癌症研究基金会/美国癌症研究所(WCRF/AICR)所推荐饮食及生活方式的评分与血浆中奇链脂肪酸、丝氨酸、甘氨酸和特定磷脂酰胆碱的水平呈正相关,在巢式病例对照研究(1 608对结直肠癌病例和相匹配对照)中该代谢谱与结直肠癌发病风险降低有关,其关联强度高于WCRF/AICR评分本身与结直肠癌的关联,提示代谢谱能够反映包含了多种行为和生物学暴露的效应,可用来更好地评估结直肠癌发病风险[28]。此外,美国范德堡大学研究团队基于中国上海女性健康研究队列和上海男性健康研究队列开展了巢式病例对照研究(250对结直肠癌病例和相匹配对照),在非靶向检测的618个血浆代谢物中,发现35个代谢产物与结直肠癌风险相关,其中12个代谢物是甘油磷脂(9个与结直肠癌风险降低相关,3个与风险增加相关),提示甘油磷脂的失调可能会增加结直肠癌的风险;此外还有9个其他脂类、7个芳香族化合物、5个有机酸和4个其他代谢物也与结直肠癌发病相关。经相互调整后,发现9个代谢物与结直肠癌独立相关,利用这些代谢物建立风险预测模型,其曲线下面积(area under curve)为0.76,相关结果尚需进一步独立验证[29]。

除上所述,其他恶性肿瘤如胰腺癌[30-32]、肝癌[33-34]、卵巢癌[35]等,国际上也有相应的代谢组流行病学研究报道,其研究设计、检测方法及分析策略等与上述研究类似,在此不再赘述。

四、代谢组流行病学研究的思考和展望

代谢组流行病学研究在反映暴露的特征和效应、揭示暴露与疾病的"黑箱"机制、发现新的生物标志物等方面开始崭露头角。然而,当前在大规模人群中开展代谢组流行病学研究仍面临需要解决

的问题。

1. 代谢物的稳定性问题：尽管代谢组测定的样品可以是血液等生物体液，对人体损伤较小，易于推广应用，但必须考虑生物标本采集（如禁食状态）和处理（如从采集到处理/冷冻的时间、离心、运输、冻融次数等）对代谢物的影响。在大规模多中心研究时，有时候很难确保每个中心严格遵守相同的采集和处理流程。有研究表明，尽管有些代谢物（如胆汁酸、维生素、嘌呤/嘧啶）受到禁食、采集血液的季节、处理时间延迟等因素的影响，但大多数血浆代谢物比较稳定[36]；在-80 ℃保存几十年的标本中也能够重复出特定饮食与代谢物的相关性[19,37]，但仍需要更大样本的验证研究，并利用重复收集的标本进行纵向数据分析，评估代谢物的长期稳定性和可靠性。

2. 代谢组检测和质量控制的问题：目前尚缺乏国际统一的代谢组学检测技术流程和质控标准，不同实验室或商业公司在标本提取、仪器配置、质控方法、代谢物资源库等方面存在较大差异，限制了跨时间、跨平台和跨研究的比较。在非靶向代谢组学研究中，获得的代谢产物水平通常是半定量浓度，而非绝对定量，导致难以确定人群实际应用的阈值，需要考虑综合运用非靶向和靶向方法。现有的代谢物数据库也在持续更新，以纳入之前未知或无法识别的代谢物。

3. 数据预处理和统计分析的问题：由于代谢组学数据的预处理步骤繁琐、代谢物数量多、相互关系复杂，增加了数据分析难度，尚无明确的最佳实践准则。尽管目前常用的算法在一定程度上提供了有效工具，但数据预处理和分析步骤可以通过不同的方法执行且没有明确的顺序，不同的基线校正、谱峰处理、变量转换、缺失填补、数据降维等方法均有可能影响最终结果，并导致在重现结果、比较结果及荟萃分析等方面出现问题。因此，亟需建立数据预处理流程、统计分析路径、可视化方案、生物学解释及报告的规范化标准协议。

4. 结果的解读与因果推断：即使是基于高质量队列的代谢组流行病学研究，仍有可能受到残余混杂或反向因果等问题的影响，导致所发现的关联并非因果关系。外部验证是评价结果真实性的重要方法，也是希尔准则（Hill's criteria）中的一致性标准所要求的——多项独立研究结果的一致性越高，因果关系的可能性就越大。由于并非所有的代谢物具有长期稳定性，外部验证尤为重要。但如果难以获取外部验证的样本（如罕见疾病），可以考虑交叉验证的方法，先对研究对象的一个子集进行主要分析，然后在剩余子集中验证结果[38]。

此外，尽管从数据中挖掘相关关系的方法研究发展迅速，但分析因果关系的方法仍十分有限。孟德尔随机化（Mendelian randomization，MR）方法的提出为基于观察性流行病学研究的因果推断提供了新路径。MR方法应用与暴露因素（如代谢物）相关联的遗传变异作为工具变量，能够克服观察性研究中的混杂和反向因果问题，为因果推断提供有力的证据[39]。随着代谢物水平相关遗传数据的累积[40]，MR方法有望在代谢组流行病学研究的因果推断中发挥重要作用。

为了加快提高代谢组学的人群研究水平，目前已建立了一些网络共享资源供参考，如代谢组学数据库（http://metabolomicssociety.org/）、生物信息学工具汇编[41]、代谢数据交换平台（http://www.metabolomexchange.org/site/）等，并成立了国际协作组。如2014年成立的代谢组研究联盟（Consortium of Metabolomics Studies）为一项大规模国际合作，包含了来自亚洲、欧洲、北美和南美的47项前瞻性队列，涉及13.6万名参与者的血液代谢组学数据，旨在通过整合资源、开发新的网络分析平台、共享源代码等方式，促进慢性疾病的代谢标志物和病因学研究[42]。2017年由复旦大学牵头启动的“国际人类表型组计划（一期）”项目，将针对以代谢组为核心的表型组（phenomics）开展流行病学研究，助力标准化流程的建立，获得反映中国人群特征的代谢组数据库。

综上所述，代谢组学技术的快速发展，给流行病学研究提供了新的机遇。在不断完善标准化体系的基础上，推动大规模的代谢组流行病学研究，将为系统流行病学研究奠定重要基础。通过整合基因组、转录组和蛋白质组，构建涵盖DNA、mRNA、蛋白质到代谢产物的调控网络，将极大加深我们对于慢性病病因及机制的理解，并有助于发现新型生物标志物，提升疾病的早期预防和干预能力。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] 黄涛，李立明. 系统流行病学[J]. 中华流行病学杂志，2018，39(5)：694-699. DOI:10.3760/cma.j.issn.0254-6450.2018.05.031.
Huang T, Li LM. Systems epidemiology[J]. Chin J Epidemiol, 2018, 39(5):694-699.DOI:10.3760/cma.j.issn.0254-6450.2018.05.031.

[2] Goodacre R, Vaidyanathan S, Dunn WB, et al. Metabolomics by numbers: acquiring and understanding global metabolite data[J]. Trends Biotechnol, 2004, 22(5): 245-252. DOI:10.1016/j.tibtech.2004.03.007.

[3] Brennan L, Hu FB. Metabolomics-based dietary biomarkers in nutritional epidemiology-current status and future opportunities[J]. Mol Nutr Food Res, 2019, 63(1):e1701064. DOI:10.1002/mnfr.201701064.

[4] Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine[J]. Nat Rev Drug Discov, 2016, 15(7):473-484. DOI:10.1038/nrd.2016.32.

[5] Emwas AH, Roy R, McKay RT, et al. NMR spectroscopy for metabolomics research[J]. Metabolites, 2019, 9(7). DOI: 10.3390/metabo9070123.

[6] Beale DJ, Pinu FR, Kouremenos KA, et al. Review of recent developments in GC-MS approaches to metabolomics-based research[J]. Metabolomics, 2018, 14(11): 152. DOI: 10.1007/s11306-018-1449-2.

[7] Gathungu RM, Kautz R, Kristal BS, et al. The integration of LC-MS and NMR for the analysis of low molecular weight trace analytes in complex matrices[J]. Mass Spectrom Rev, 2020, 39(1-2):35-54. DOI:10.1002/mas.21575.

[8] Dunn WB, Broadhurst D, Begley P, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry[J]. Nat Protoc, 2011, 6(7): 1060-1083. DOI:10.1038/nprot.2011.335.

[9] Playdon MC, Joshi AD, Tabung FK, et al. Metabolomics analytics workflow for epidemiological research: Perspectives from the Consortium of Metabolomics Studies (COMETS) [J]. Metabolites, 2019, 9(7): 145. DOI: 10. 3390/metabo9070145.

[10] Gromski PS, Xu Y, Kotze HL, et al. Influence of missing values substitutes on multivariate analysis of metabolomics data[J]. Metabolites, 2014, 4(2): 433-452. DOI:10.3390/metabo4020433.

[11] Kokla M, Virtanen J, Kolehmainen M, et al. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study [J]. BMC Bioinformatics, 2019, 20(1): 492. DOI: 10.1186/s12859-019-3110-0.

[12] Shah J, Brock GN, Gaskins J. BayesMetab: treatment of missing values in metabolomic studies using a Bayesian modeling approach[J]. BMC Bioinformatics, 2019, 20 Suppl 24:673. DOI:10.1186/s12859-019-3250-2.

[13] Faquih T, van Smeden M, Luo J, et al. A workflow for missing values imputation of untargeted metabolomics data[J]. Metabolites, 2020, 10(12): 486. DOI: 10.3390/metabo 10120486.

[14] Westfall P, Young S. Resampling-based multiple testing: examples and methods for p-value adjustment[M]. New York:John Wiley & Sons, 1993.

[15] Zhou ZR, Wang WW, Li Y, et al. In-depth mining of clinical data: the construction of clinical prediction model with R [J]. Ann Transl Med, 2019, 7(23):796. DOI:10.21037/atm.2019.08.63.

[16] Sun Y, Gao HY, Fan ZY, et al. Metabolomics signatures in type 2 diabetes: a systematic review and integrative analysis[J]. J Clin Endocrinol Metab, 2020, 105(4):dgz240. DOI:10.1210/clinem/dgz240.

[17] Merino J, Leong A, Liu CT, et al. Metabolomics insights into early type 2 diabetes pathogenesis and detection in individuals with normal fasting glucose[J]. Diabetologia, 2018, 61(6):1315-1324. DOI:10.1007/s00125-018-4599-x.

[18] Qiu G, Zheng Y, Wang H, et al. Plasma metabolomics identified novel metabolites associated with risk of type 2 diabetes in two prospective cohorts of Chinese adults[J]. Int J Epidemiol, 2016, 45(5):1507-1516. DOI:10.1093/ije/dyw221.

[19] Hang D, Zeleznik OA, He X, et al. Metabolomic signatures of long-term coffee consumption and risk of type 2 diabetes in women[J]. Diabetes Care, 2020, 43(10): 2588-2596. DOI:10.2337/dc20-0800.

[20] Li J, Guasch-Ferre M, Chung W, et al. The Mediterranean diet, plasma metabolome, and cardiovascular disease risk [J]. Eur Heart J, 2020, 41(28): 2645-2656. DOI: 10.1093/eurheartj/ehaa209.

[21] Holmes MV, Millwood IY, Kartsonaki C, et al. Lipids, lipoproteins, and metabolites and risk of myocardial infarction and stroke[J]. J Am Coll Cardiol, 2018, 71(6): 620-632. DOI:10.1016/j.jacc.2017.12.006.

[22] Seow WJ, Shu XO, Nicholson JK, et al. Association of untargeted urinary metabolomics and lung cancer risk among never-smoking women in China[J]. JAMA Netw Open, 2019, 2(9):e1911970. DOI:10.1001/jamanetworkopen.2019.11970.

[23] Wen CP, Zhang F, Liang D, et al. The ability of bilirubin in identifying smokers with higher risk of lung cancer: a large cohort study in conjunction with global metabolomic profiling[J]. Clin Cancer Res, 2015, 21(1): 193-200. DOI:10.1158/1078-0432.CCR-14-0748.

[24] His M, Viallon V, Dossus L, et al. Prospective analysis of circulating metabolites and breast cancer in EPIC[J]. BMC Med, 2019, 17(1):178-178. DOI:10.1186/s12916-019-1408-4.

[25] Moore SC, Playdon MC, Sampson JN, et al. A metabolomics analysis of body mass index and postmenopausal breast cancer risk[J]. J National Cancer Inst, 2018, 110(6): 588-597. DOI:10.1093/jnci/djx244.

[26] Moore SC, Mazzilli KM, Sampson JN, et al. A metabolomics analysis of postmenopausal breast cancer risk in the Cancer Prevention Study Ⅱ [J]. Metabolites, 2021, 11(2). DOI:10.3390/metabo11020095.

[27] Tian J, Xue W, Yin H, et al. Differential metabolic alterations and biomarkers between gastric cancer and colorectal cancer: a systematic review and Meta-analysis [J]. Onco Targets Ther, 2020, 13:6093-6108. DOI:10.2147/OTT.S247393.

[28] Rothwell JA, Murphy N, Besevic J, et al. Metabolic signatures of healthy lifestyle patterns and colorectal cancer risk in a european cohort[J]. Clin Gastroenterol Hepatol, 2020, 29: S1542-3565. DOI: 10.1016/j. cgh. 2020. 11.045.

[29] Shu X, Xiang YB, Rothman N, et al. Prospective study of blood metabolites associated with colorectal cancer risk [J]. Int J Cancer, 2018, 143(3): 527-534. DOI: 10.1002/ijc.31341.

[30] Shu X, Zheng W, Yu D, et al. Prospective metabolomics study identifies potential novel blood metabolites associated with pancreatic cancer risk[J]. Int J Cancer, 2018, 143(9):2161-2167. DOI:10.1002/ijc.31574.

[31] Fest J, Vijfhuizen LS, Goeman JJ, et al. Search for early pancreatic cancer blood biomarkers in five european prospective population biobanks using metabolomics[J]. Endocrinology, 2019, 160(7): 1731-1742. DOI: 10.1210/en.2019-00165.

[32] Adam MG, Beyer G, Christiansen N, et al. Identification and validation of a multivariable prediction model based on blood plasma and serum metabolomics for the distinction of chronic pancreatitis subjects from non-pancreas disease control subjects[J]. Gut, 2021. DOI: 10.1136/gutjnl-2020-320723.

[33] Assi N, Thomas DC, Leitzmann M, et al. Are metabolic signatures mediating the relationship between lifestyle factors and hepatocellular carcinoma risk? Results from a nested case-control study in EPIC[J]. Cancer Epidemiol Biomarkers Prev, 2018, 27(5): 531-540. DOI: 10.1158/1055-9965.EPI-17-0649.

[34] Loftfield E, Rothwell JA, Sinha R, et al. Prospective investigation of serum metabolites, coffee drinking, liver cancer incidence, and liver disease mortality[J]. J Natl Cancer Inst, 2020, 112(3): 286-294. DOI: 10.1093/jnci/djz122.

[35] Zeleznik OA, Eliassen AH, Kraft P, et al. A prospective analysis of circulating plasma metabolites associated with ovarian cancer risk[J]. Cancer Res, 2020, 80(6): 1357-1367. DOI:10.1158/0008-5472.CAN-19-2567.

[36] Townsend MK, Clish CB, Kraft P, et al. Reproducibility of metabolomic profiles among men and women in 2 large cohort studies[J]. Clin Chem, 2013, 59(11): 1657-1667. DOI:10.1373/clinchem.2012.199133.

[37] Zheng Y, Yu B, Alexander D, et al. Human metabolome associates with dietary intake habits among African Americans in the atherosclerosis risk in communities study[J]. Am J Epidemiol, 2014, 179(12): 1424-1433. DOI: 10.1093/aje/kwu073.

[38] Triba MN, Le Moyec L, Amathieu R, et al. PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters[J]. Mol Biosyst, 2015, 11(1): 13-19. DOI: 10.1039/c4 mb00414k.

[39] Gala H, Tomlinson I. The use of Mendelian randomisation to identify causal cancer risk factors: promise and limitations[J]. J Pathol, 2020, 250(5): 541-554. DOI: 10.1002/path.5421.

[40] Hagenbeek FA, Pool R, van Dongen J, et al. Author Correction: Heritability estimates for 361 blood metabolites across 40 genome-wide association studies [J]. Nat Commun, 2020, 11(1): 1702. DOI: 10.1038/s41467-020-15276-y.

[41] Misra BB, Mohapatra S. Tools and resources for metabolomics research community: A 2017-2018 update [J]. Electrophoresis, 2019, 40(2): 227-246. DOI: 10.1002/elps.201800428.

[42] Yu B, Zanetti KA, Temprosa M, et al. The Consortium of Metabolomics Studies (COMETS): metabolomics in 47 prospective cohort studies[J]. Am J Epidemiol, 2019, 188(6):991-1012. DOI:10.1093/aje/kwz028.