



化学计量学在罕见病与常见复杂疾病代谢组学研究中的应用

鲁艳^{1#}, 肖雯^{1#}, 惠羽凡¹, 杨琴^{2*}, 吴莉萍¹, 杨江涛¹, 姜盼盼^{1*}

(1. 深圳爱湾医学检验实验室 深圳市罕见病代谢组学精准医学工程研究中心, 广东 深圳 518000;
2. 长江大学 物理与光电工程学院, 湖北 荆州 434023)

摘要: 罕见病和常见复杂疾病的研究至关重要, 这些疾病影响着大量患者的生活质量, 且往往面临诊断难度和治疗方案的局限性。罕见疾病通常由于其低发病率和复杂的遗传基础, 使得早期诊断和有效治疗变得更加困难。而常见复杂疾病因其多因素致病特性, 导致预防、诊断和治疗面临挑战。组学技术, 尤其是代谢组学, 提供了对生物系统中代谢物的全面评估, 能够揭示疾病进展、筛选潜在生物标志物并提供病理生理学的深刻见解。为了解释组学分析所生成的庞大数据, 化学计量学作为一个跨学科领域, 结合了计算机科学、数学和统计学, 展现出强大的数据处理能力。该文综述了化学计量学在罕见疾病和常见复杂疾病组学研究中的应用, 重点分析其在生物标志物筛选、疾病诊断和个性化治疗中的重要性, 并探讨了机器学习在数据分析中的角色。总体而言, 化学计量学为深入理解疾病机制和推动医疗创新提供了重要支持。

关键词: 化学计量学; 代谢组学; 遗传代谢病; 泌尿系结石; 糖尿病

中图分类号: O657.63; TB9 **文献标识码:** A **文章编号:** 1004-4957(2025)12-2629-12

Application of Chemometric Techniques in Metabolomic Research on Rare and Common Complex Diseases

LU Yan^{1#}, XIAO Wen^{1#}, HUI Yu-fan¹, YANG Qin^{2*}, WU Li-ping¹,
YANG Jiang-tao¹, JIANG Pan-pan^{1*}

(1. Shenzhen Rare Disease Engineering Research Center of Metabolomics in Precision Medicine, Shenzhen Aone Medical Laboratory Co., Ltd, Shenzhen 518000, China; 2. School of Physics and Optoelectronic Engineering, Yangtze University, Jingzhou 434023, China)

Abstract: Research on rare diseases and common complex diseases is crucial, as these conditions significantly impact the quality of life for many patients and often present challenges in diagnosis and treatment. Rare diseases are typically more challenging for early diagnosis and effective treatment due to their low incidence rates and complex genetic basis. Common complex diseases, characterized by multifactorial etiologies, pose significant challenges for prevention, diagnosis, and treatment. Omics technologies, particularly metabolomics, offer a comprehensive assessment of metabolites within biological systems, revealing disease progression, screening potential biomarkers, and providing profound insights into pathophysiology. To interpret the vast data generated by omics analyses, chemometrics, an interdisciplinary field combining computer science, mathematics, and statistics, demonstrates powerful data processing capabilities. This review discusses the application of chemometrics in the omics research of rare and common complex diseases, highlighting its significance in biomarker discovery, disease diagnosis, and personalized treatment, while also exploring the role of machine learning in data analysis. Overall, chemometrics provides essential support for a deeper

收稿日期: 2025-02-12; 修回日期: 2025-03-29

基金项目: 深圳市工程研究中心(工程实验室)组建项目(F-2020-Z99-502615)

* 通讯作者: 杨琴, 博士, 副教授, 研究方向: 化学计量学及其在疾病检测和机理分析中的应用研究, E-mail: yangqin00055@hnu.edu.cn
姜盼盼, 硕士, 助理研究员, 研究方向: 化学计量学以及多组学数据分析在疾病检测中的应用研究, E-mail: jiangpanpan@aonebio.com.cn

共同第一作者

网络首发时间: 2025-09-11

understanding of disease mechanisms and the advancement of medical innovation.

Key words: chemometrics; metabolomics; inherited metabolic disease; urolithiasis; diabetes

化学计量学(Chemometrics)这一概念由 Svante Wold 于 1971 年首次提出, 1974 年在 Bruce R. Kowalski 与 Svante Wold 的共同倡议下, 成立了化学计量学学会^[1]。以往分析仪器所获得的相关数据非常有限, 但随着科技的进步、新型仪器和分析方法的引入, 现代分析仪器提供了大量可靠的测量数据^[2]。目前, 各类分析化学研究对象的数据大多具有构成组分种类和数量多样化、相互之间协同关系复杂化与非线性化的特点。这种庞大而复杂的高维数据, 如果没有适当工具, 难以对其进行描述与分析^[3]。传统的单变量分析方法主要研究一个变量对研究对象的影响或者关联程度, 无法考虑多个变量之间的关系和相互作用, 不能判断相关或冗余特征。化学计量学跨学科融合计算机科学、数学和统计学, 旨在从复杂的化学测量数据中最大限度地提取有价值的信息^[4]。这种多变量方法能够同时综合处理多个变量, 获得变量间的联系和数据结构等信息, 对复杂体系样本的内在规律和关键变量进行综合解析。化学计量学发展迅速, 各种新的算法及相关应用取得重要进展, 构建了从数据预处理、特征提取到模型构建与验证的完整分析流程, 成为化学与分析化学领域的重要研究方向, 在药物研发、食品科学、环境检测、临床诊断等领域都有重要应用。

化学计量学强大的数据处理与分析能力不仅革新了化学研究的范式, 还为其他相关领域, 特别是在代谢组学研究中, 开辟了新的研究路径。化学计量学方法通过比较健康与患病状态下的组学数据, 能够揭示潜在的疾病机制, 有助于理解复杂的生物学过程, 以及如何通过调节特定的代谢通路或蛋白质相互作用来影响疾病的进程。代谢物是生化途径的下游产物, 能够显示基因型和表型之间的联系, 反映基因组、表观基因组和环境变化的影响^[5]。由于代谢组学能够对多种因素引起的生物体变化做出反应, 因此它是跟踪、监测和了解这些影响与疾病之间关系的理想方法^[6-7]。通过代谢组学技术, 使用色谱(高效液相色谱(HPLC)或气相色谱(GC)结合质谱(MS)能够对生物样本(如血液、尿液等)中的小分子代谢产物进行高通量、高灵敏度的分析, 全面反映生物体的代谢状态, 揭示与疾病相关的代谢变化^[8-9]。基于化学计量学, 代谢组学充分利用复杂的化学测量数据, 揭示生物体内代谢活动的深层规律。通过建立多变量模型, 深入挖掘数据中的隐藏信息, 为高维数据的整合和分析提供了强有力的支持, 为数据的高效处理和分析提供了可能。化学计量学在疾病与组学研究中的应用, 不仅为理解生物系统的复杂性提供了工具, 也推动了疾病的早期诊断、治疗靶点的发现和个性化医疗的发展。

化学计量学通过数学方法对组学数据中的复杂混合信息进行深入分析, 提取与研究对象内涵与本质规律紧密相关的信息。在化学计量学领域, 机器学习作为一种有效的工具, 得到了广泛应用, 尤其在方法开发与数据分析方面, 其处理和分析复杂实验数据的能力尤为显著^[10]。机器学习能够自动从数据中学习规律和模式, 根据问题类型(分类、回归、聚类等)选择最优学习算法从训练数据中提取出有用的特征并优化模型参数, 通过不断调整模型结构以最小化预测误差, 逐渐提升对新数据的预测和分类能力。通常可以将机器学习分为监督学习和无监督学习。监督学习通过分析自变量(特征)和因变量(标签/类别)之间的关系, 从特征中预测新数据的标签。常见的监督学习算法有线性判别分析(LDA)^[11]、偏最小二乘(PLS)^[12]、支持向量机(SVM)^[13]等。无监督学习用于处理未标记的数据, 找到将数据聚类的固有结构。常见的无监督学习算法有聚类分析和主成分分析(PCA)^[14]。此外, 深度学习作为机器学习的一个分支, 通过使用深度神经网络来学习数据的表示和提取特征。

罕见疾病由于其低发病率、复杂的遗传基础以及多样的临床表现, 通常面临诊断难题和治疗手段的局限性。而复杂疾病, 由于其多因素致病特性, 导致疾病的预防、诊断和治疗面临多重挑战。本综述简要阐述了罕见疾病和常见复杂疾病存在的问题, 深入探讨并全面综述了化学计量学在罕见疾病及常见复杂疾病组学数据分析中的应用(图 1), 并讨论了化学计量学方法体系及其优化算法等如何应用于疾病标志物的筛选、诊断、治疗和预后。

1 化学计量学在罕见病中的应用

罕见病是指发生在极少数人身上的罕见疾病, 目前尚无统一的或被广泛接受的定义。为促进跨国界的交流、知识共享及孤儿药的协调开发, 世界卫生组织(WHO)将罕见病定义为发病率在 0.065%~0.1%之间的疾病或病变。其中, 约 80%的罕见病是遗传缺陷引起的, 常引发严重的生理功能障碍, 显

著缩短预期寿命, 并对身心健康造成损害^[15]; 仅有 5% 的罕见病可以得到有效的干预或治疗; 约 50% 左右的罕见病在新生儿阶段或儿童时期开始发病, 如先天性代谢缺陷或遗传代谢病, 这类疾病通常源于遗传因素且发病较早, 贯穿患者生命的大部分时间, 进一步加重了疾病负担^[16]。罕见病不仅构成了重要的公共卫生问题, 也给医学界带来了诸多挑战^[15]。由于罕见疾病的潜在生物学机制非常复杂, 开展相关研究和制定有效治疗策略面临困难。医生在识别和诊断罕见疾病方面的经验通常有限, 导致患者获得准确诊断的过程复杂且漫长。此外, 罕见疾病患者数量较少, 招募及开展临床研究也面临诸多限制^[17]。罕见疾病研究是“精准医学”的重要组成部分, 对于许多罕见疾病而言, 早期诊断和治疗能够显著改善患者的预后和生活质量。在此背景下, 化学计量学作为一种强大的工具, 已广泛应用于疾病的早期检测与辅助诊断。该方法利用蛋白质组学、基因组学和代谢组学指纹图谱进行建模, 能够实现对定量属性(如蛋白质含量的连续值)和定性属性(如健康/疾病的离散状态)的预测^[18]。通过这一过程, 化学计量学可有效地将数据内在信息转化为可理解的知识, 为罕见病的深入研究提供了有力的支持。

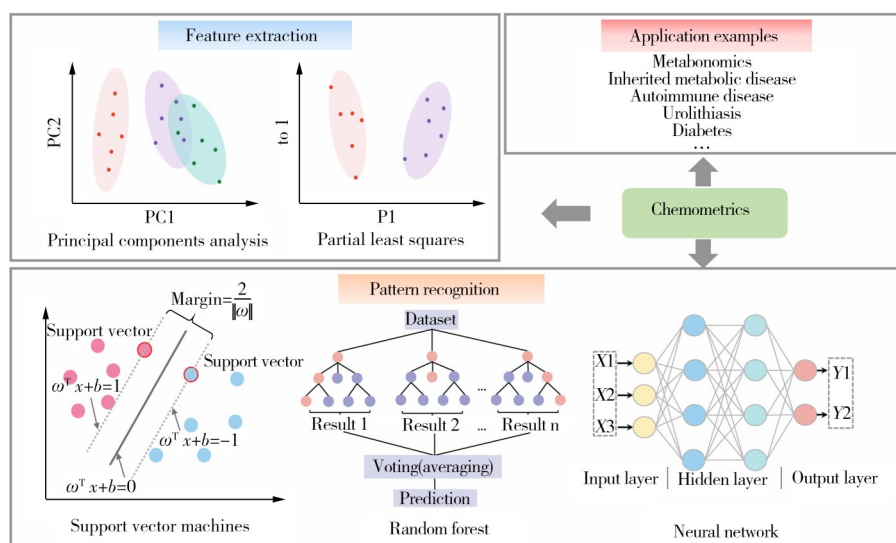


图 1 化学计量学研究方法及应用领域
Fig. 1 Research methods and applications in chemometrics

1.1 遗传代谢病

新生儿遗传代谢病(IMDs)也称先天性代谢异常, 是指具有异常生化代谢标志物的一大类疾病。这类疾病因基因突变引起酶活性改变、蛋白质结构或功能改变, 导致机体内部代谢紊乱, 有毒中间代谢物异常蓄积, 引起轻重不等的一系列临床症状, 并产生严重病理后果, 严重时可能会导致终身残疾或死亡^[19]。这类疾病临床症状无特异性, 进行快速识别和特征标记物筛选是实现早期筛查和诊断的关键所在。当前分析检测技术的快速发展使分析所得到的数据具有很高的空间复杂程度(如色谱、光谱以及高通量质谱检测技术得到的数据通常具有很高的维数)。

1.1.1 遗传代谢病的化学计量学方法应用 代谢组学结合化学计量学为构建IMDs病理图谱和快速识别模型提供了全新路径, 能够对高维数据进行特征选择和特征提取, 降低数据维数。PCA是一种无监督学习算法, 将原始变量空间转换到低维子空间, 通过寻找原始变量的线性组合来最大化数据方差^[20]。因此, PCA能够在信息损失最小的情况下降低数据维度。Moreda-Piñeiro等^[21]对健康新生儿和患有IMDs的新生儿的干血斑数据进行化学计量学分析, PCA和聚类分析显示, 样本有明显的分组趋势, 进一步使用LDA和簇类独立软模式法(SIMCA)建模, 正确率分别为94.05%和86.37%。LDA是一种有监督学习算法, 将数据投影到低维空间时可尽可能地保留数据样本信息, 使不同类别的样本尽可能地分开, 同一类别内的样本尽可能地接近。Zaunseder等^[22]使用包含多变量和标签的异戊酸尿症(IVA)数据来训练LDA, 并将特征转换为LDA维数, 作为训练和评估逻辑回归(LR)分类器的输入。然后将训练好的LDA和LR模型应用于测试数据集的变量, 并将预测结果与测试数据标签进行比较, 以评估模型。

偏最小二乘判别分析(PLS-DA)不同于PCA, 是一种有监督学习算法, 结合了PLS和判别分析的思想。该算法通过投影和降维技术, 深入挖掘数据的内在结构, 在数据的特征空间中寻找能够最大化区

分不同类别样本的潜在变量,实现对新样本的准确分类。与其他多元线性模型相比,PLS能够在自变量存在严重多重相关性、样本点少于变量个数的条件下建模。Liu等^[23]使用Orbitrap质谱分析仪对20例IMDs患者和20例健康儿童的尿液样本进行了基于非靶向代谢组学的PLS-DA分析,发现两者之间存在明显的分离趋势。此外,PLS-DA在数据分析中展现出卓越的数据解释能力,尤其在处理高维数据和复杂模式识别任务时具有独特的优势。它通过信息向量从不同角度度量变量的重要性,能够准确地识别出对分类最为关键的变量。通过揭示哪些变量(如代谢物)在区分不同类别(如健康与疾病状态)中起到关键作用,PLS-DA为生物医学研究提供了宝贵的线索和启示。Kalivodová等^[24]将PLS-DA用于23例中链酰基辅酶A脱氢酶缺乏症(MCADD)和23例健康对照组的血液样本数据分析,实现了疾病识别和标志物筛选,并筛选出辛酰肉碱(C8)、己酰肉碱(C6)、癸酰肉碱(C10)和癸烯酰肉碱(C10:1)等疾病标志物。作为PLS-DA模型的扩展,正交偏最小二乘判别分析(OPLS-DA)同样能够通过隐变量提取的方法解决数据中的共线性和信息冗余问题^[25]。Weerd等^[26]比较了苯丙酮尿症(PKU)患者和对照组的非靶向脂质组学(22 PKU, 22 对照)和代谢组学(35 PKU, 20 对照)的数据。OPLS-DA得分图显示各组之间存在明显差异。PKU患者的苯丙氨酸、苯丙氨酸代谢物、苯丙氨酸结合物以及含苯丙氨酸的二肽显著升高。此外,与健康个体相比,PKU患者表现出不同的脂质特征,其中差异最显著的脂质分子属于甘油磷脂和甘油酯类。

1.1.2 化学计量学算法的改进与优化策略 由于代谢组学数据集规模和复杂性的增加,化学计量学方法的性能常常会下降。全局优化技术和集成学习策略对传统化学计量学方法性能的完善能有效解决疾病识别和特征标记物筛选难题,作为模型结构参数调节和模型组合技术,在改善模型性能方面显示出独特优势。全局优化技术(如粒子群优化算法(PSO)),通过群体协作和信息共享快速找到最优解,实现多参数协同优化,提升算法泛化性能。本课题组^[27]采用混合粒子群优化(HPSO)结合OPLS-DA,同时选择最优变量子集、相关权重和正交成分数量,构建新算法HPSO-OPLSDA。对甲基丙二酸血症(MMA)和IVA的研究结果表明,该算法能够显著提高OPLS-DA在IMDs检测中的性能,并识别出已报道的诊断性生物标志物。Chen等^[28]结合离散粒子群优化(DPSO)和堆叠自动编码器(SAE)形成DPSO-SAE集成框架,用于诊断两种遗传代谢病(健康 *vs* 戊二酸血症I型(GA-I)、丙酸血症(PA))。SAE具有出色的非线性特征学习能力,DPSO的引入使SAE能够发现生物标志物,并通过协同优化变量组合和SAE建模中神经元的数量,提高模型的性能。此外,本课题组^[29]还开发了一种基于穷尽划分和平行排列的半对半分分解策略(EPHAH),结合OPLS-DA建立多类别IMDs识别,EPHAH能够在避免累积误差的基础上集成所有OPLS-DA二元分类器的模式识别能力,充分学习多类问题的复杂决策边界(PA、3-甲基巴豆酰辅酶A缺乏症(MCCD)、MMA、酪氨酸血症(TYR)、GA-I、PKU、IVA、鸟氨酸氨甲酰基转移酶缺乏症(OTCD)、枫糖尿症(MSUD)),提高分类性能。

单个模型在选择稳定性和可重复性方面存在问题。集成学习策略基于集体的智慧,通过多模型集体决策机制,利用多个模型的优势,可有效克服单一模型固有的偏差和局限性,显著提高预测结果的准确性和可靠性,从而实现更好的泛化能力和准确性。Zhu等^[30]开发了基于机器学习的多变量筛选模型,用作降低MMA假阳性的二级诊断工具,有助于减轻患者及其家属在MMA诊断过程中的情感和经济负担,以及公共卫生系统的负担。该模型使用随机森林(RF)、C5.0、稀疏线性判别分析和以算法随机梯度提升为监督器的自编码器深度神经网络组合,取得了最佳预测结果。此外,单个模型在小样本下筛选的特征(生物标志物)可能不稳定,易受样本变化影响,多个模型即使性能相似,选出的特征也可能不同。集成学习策略能够有效开发和综合利用多个单个模型的特征变量选择能力,提高模型筛选稳健性。本课题组^[31]将自助法(Bootstrap)与基于协同自组织映射(SOM)的OPLS-DA相结合(BS-sSOM-OPLSDA)用于诊断MMA和PA。Bootstrap用于生成多个扰动数据集并建模,通过整合多个模型选择结果的输出排名(变量投影重要性VIP),筛选出在多个数据集中一致被识别为重要的变量。结果表明,所提出的特征筛选系统在分类性能、识别标志物与代谢机制的相关性以及所选标志物对样本变化的稳定性方面,都优于更简单的技术。随后本课题组^[32]还开发了一种新混合扰动框架BS-MIVF-PLSDA对每个重抽样构成的扰动训练样本,在建模算法层面进一步采用多信息向量融合策略(MIVF)进行扰动,用以获取在变量空间更强的搜索能力,进而筛选出性能更优越的稳健生物标志物,进一步扩展集成策略对提高变量筛选稳健性的优势。

采用化学计量学方法解析多组学融合数据, 提取有意义的生物信息解决各种临床医学问题已成为分析化学与生命科学交叉领域的一个重要方向。但是, 由于来自不同组学的融合数据呈现高维、高噪、小样本以及数据信息不平衡和非线性等特性, 传统化学计量学方法存在一定的局限性。多块数据融合策略考虑了多组学数据集之间的异质性, 同时对多个数据矩阵建模, 从多个数据集中提取特征, 充分考虑组学内和组学间的关系, 能够得到更可靠的分析结果。建模多基于已建立的多变量方法(如: PLS)进行扩展。本课题组^[33]基于尿液和血液代谢特征对疾病病理的联合表征, 提出了一种新的中级组学数据融合策略, 用于 MMA 亚型的有效分类。该策略不仅考虑了单个数据块内的变量关系, 而且通过使用 PCA 的主成分(PC)导向 PLS 探索数据块之间的偏相关性, 扩大多变量优势。此外, 在 Bootstrap 的基础上, 开发了集成特征筛选框架来筛选稳定的特征变量。其他方法如 O2PLS^[34]和 OnPLS^[35]被用于融合两个及以上组学数据集, OnPLS 将来自多个组学数据矩阵的数据结构分为全局、局部多元数据间关系和单个多元数据内在关系^[36]。block PLSDA^[37]则用于将多组学数据投影到潜变量空间后进行判别分析建模。本课题组^[38]采用 block PLSDA 融合代谢组学和蛋白质组学, 探索钴胺素 C(cblC)缺乏症 3 种临床表型(癫痫、发育迟缓和代谢综合征)之间的差异。

1.1.3 机器学习与深度学习在遗传代谢病中的应用进展 机器学习和深度学习也逐渐被应用于遗传代谢病领域。Usha Rani 等^[39]使用机器学习工具整合多种 IMDs 人口统计学数据、代谢组学数据和基因组学数据, 建立特定疾病的切值, 进一步确定主要和次要标志物, 并构建分类和回归树(CART)模型以进行鉴别诊断。Leal-Witt 等^[40]为了解决 IMDs 中小样本量和类不平衡问题, 采用 XGBoost(eXtreme Gradient Boosting)算法完成学习器间的迭代, 并进行 Ridge 正则化(L2), 对多个 XGBoost 参数进行超参数调整, 特征重要性分析使用 SHAP(SHapley Additive exPlanations)为模型的每个特征分配重要性值, 解释模型的预测过程。Li 等^[41]在大型人群筛查数据集(355 197 个样本)上训练机器学习模型辅助 IMDs 临床诊断或筛查, 但由于 IMDs 合计仅占有所有病例的 3.3%, 并且各种 IMDs 之间也存在显著差异, 因此该课题组结合加权欠采样或过采样策略以及两阶段 RF 建模来应对分类不平衡问题, 以达到罕见病筛查所需的高灵敏度。Groen 等^[42]收集了 3 958 个样本(其中 31 名患有糖原累积病 Ia(GSD Ia)), 通过支持向量机合成少数过采样技术(SVMSMOTE)生成合成样本和随机欠采样以解决数据集中的类别不平衡问题, 并基于梯度提升树构建了多个机器学习模型(RF、XGBoost 和 CatBoost), 该模型表现出高灵敏度和特异性, 所选参数不仅稳健且高度可解释。Messa 等^[43]为了弥合真实代谢特征和模拟代谢特征之间的差距, 开发了一种基于 Siamese 神经网络(SNN)的方法, 针对来自 IMDs 患者的相对较小的代谢谱数据集, 可以提供更多没有患者数据的 IMDs 信息, 从而生成模拟代谢谱, 并且能够利用代谢网络内高度复杂的相互作用, 识别未知的疾病标志物或它们的组合。

1.2 自身免疫性疾病

自身免疫性疾病(Autoimmune disease)是在没有持续感染或其他可识别诱因情况下, 由 T 细胞、B 细胞或两者共同激活引发的临床综合征, 涉及多个器官的各种免疫功能障碍^[44]。除类风湿性关节炎和自身免疫性甲状腺炎等外, 部分自身免疫性疾病较罕见, 其中自身免疫性脑炎(AE)、自身免疫性垂体炎(AH)和自身免疫性胰岛素受体病(AIR)被纳入我国罕见病目录^[45]。尽管分子、免疫学、遗传学和临床研究提供了对某些自身免疫性疾病机制的理解, 但对人体自身免疫性疾病的驱动因素(包括环境触发因素)和发病机制仍然知之甚少^[46]。通过代谢组学与化学计量学相结合, 可为识别该类疾病的潜在生物标志物和受干扰代谢途径以及疾病诊断和分期提供临床价值。

AE 是一种具有严重临床发作的非感染性急症, 患病比例占脑炎病例的 10%~20%, 估算年发病率为 1/10 万左右。近年来, AE 的诊断及预后风险受到越来越多的关注。Xiong 等^[47]将 NMR 代谢组学与稳健 OPLS-DA 相结合进行 AE 和耐药性癫痫(DRE)的区分以及针对 LGII 自身抗体相关的 3 种常见 AE 亚型, 鉴定特异性代谢物, 并采用多变量 ROC 分析加 LR 评估代谢物的分类潜力。该研究基于化学计量学方法提出了一种新型辅助诊断手段, 具有检测快速、成本效益高及诊断准确性优良等特点。尽管目前自身抗体检测仍被视为 AE 诊断的金标准, 但该方法在 AE 诊断及亚型鉴别方面展现出独特的临床应用价值, 为临床实践提供了重要的补充诊断工具。Guo 等^[48]利用决策树(DT)、RF、XGBoost、k 近邻(KNN)、SVM 等 9 种算法探究影响 AE 预后的因素, 构建可视化和可解释的预测模型。Zhang 等^[49]在多

中心队列研究中使用 6 种机器学习算法预测 AE 患者的长期认知结果, 共纳入 341 名患者。使用 LASSO 回归进行变量选择, 6 个变量作为预测因子, 其中 RF 模型是预测认知障碍的最有效模型, 此外, SHAP 方法的利用有助于更深入地了解各个特征的贡献和相对重要性, 为模型提供解释, 增强了其对临床医生的可用性。AH 是一类自身免疫介导的炎症侵犯下丘脑、垂体及其邻近器官的罕见疾病。机器学习算法可根据体外研究数据帮助医生确定新标志物的存在, 以及对可疑垂体病变进行无创诊断。在接受免疫检查点阻泄剂(ICB)治疗的癌症患者中, 常出现 ICB 诱导的自身免疫性垂体炎(HP), Mekki 等^[50]开发了一种基于机器学习的成像决策算法(RF)用于区分垂体转移(PM)和 HP, 并在外部验证集中达到高性能, 基于影像学的决策算法可以帮助放射科医生和肿瘤学家在接受免疫治疗的患者中对垂体异常的早期和优化管理做出正确的诊断。

表 1 列举了化学计量学在罕见病早期诊断/预测中的应用情况。

表 1 化学计量学在罕见病早期诊断/预测中的应用

Table 1 Applications of chemometrics in early diagnosis/prediction of rare diseases

Diseases	Objective	Method	Sample size	Reference
IMDs(12 种)	识别健康新生儿和患有代谢紊乱的新生儿	PCA、LDA、SIMCA	42 例健康对照、42 例代谢紊乱	[21]
IVA	改进 IVA 新生儿筛查, 保持敏感性, 降低假阳性, 区分和预测 IVA 病程(轻度与典型)	PCA、LDA、LR	2 106 090 例(其中 103 假阳、28 例 IVA)	[22]
IMDs	识别健康儿童和 IMDs 患儿, 发现新的潜在生物标志物	PLS-DA	20 例健康对照、20 例 IMDs 患儿	[23]
PKU、MCADD	识别健康儿童和 IMDs 患儿, 验证筛选出的代谢物的准确性	PLS-DA	6 例对照、6 例 PKU; 23 例对照、23 例 MCADD	[24]
PKU	通过比较 PKU 患者与对照组, 寻找新的生物标志物	PCA、OPLS-DA	脂质组学(22 例 PKU, 22 例对照)和代谢组学(35 例 PKU, 20 例对照)	[26]
MMA、IVA	开发新的算法 HPSO-OPLSDA 改善 OPLS-DA 的局限性, 提高 IMDs 的诊断能力	HPSO-OPLSDA	20 例健康对照、25 例 MMA、22 例 IVA	[27]
GA-I、PA	开发 DPSO-SAE 用于 IMDs 的早期诊断, 提高模型的分性能	DPSO-SAE	30 例健康对照、30 例 GA-I、30 例 PA	[28]
IMDs(9 种)	开发 EPHAH 结合 OPLS-DA 算法, 建立多类别 IMDs 识别方法	EPHAH、OPLS-DA	38 例健康对照、34 例 PA、33 例 MCCD、40 例 MMA、40 例 TYR、30 例 GA-I、40 例 PKU、25 例 IVA、21 例 OTCD、18 例 MSUD	[29]
MMA	改进分类模型, 降低 MMA 诊断假阳性	结合 RF、C5.0 等集成模型	600 例 MMA 假阳、33 例 MMA 真阳	[30]
MMA、PA	开发 BS-sSOM-OPLSDA 筛选具有生物学意义的生物标志物	BS-sSOM-OPLSDA	45 例 MMA、40 例 PA、42 例对照	[31]
MMA、PA	开发 BS-MIVF-PLSDA 提高生物标志物筛选稳健性	BS-MIVF-PLSDA	45 例 MMA、40 例 PA、42 例对照	[32]
MMA	开发 Novel mid-level fusion 策略, 用于 MMA 亚型分类	Novel mid-level fusion	33 例 mut、33 例 cbIC	[33]
MMA 亚型	开发一种新的计算框架来系统表征和预测 cbIC 缺陷的临床表型	block PLSDA	230 例 cbIC 临床表型(癫痫、发育迟缓、代谢综合征)、68 例对照	[38]
23 种 IMDs	建立串联质谱法中分析物特异性临界值, 降低假阳性和假阴性	CART	3 000 例新生儿(其中 99 例 IMDs)	[39]
PKU	研究 Phe 浓度与胰岛素抵抗的关系, 发现潜在生物标志物	XGBoost	24 例 PKU、24 例对照	[40]
IMDs	提高 IMDs 临床诊断或筛查灵敏度	过采样/欠采样、RF	355 197 例(33.3%IMDs)	[41]
GSD Ia	根据血酰基肉碱谱识别 GSD Ia 患者	SVM、SMOTE、RF、XGBoost、CatBoost	3 958 例(31 例 GSD Ia)	[42]
IMDs	在模拟和真实代谢数据之间建立映射, 辅助诊断 IMDs	SNN	168 例受试者(收集于其他研究)	[43]
AE、DRE	识别 AE 与 DRE, 进一步区分 AE 亚型	OPLS-DA	162 例 DRE、76 例 AE	[47]
AE	预测抗体阳性 AE 患者的预后	DT、RF、XGBoost 等	187 例 AE	[48]
AE	预测 AE 患者长期认知结局	MLP、KNN、LR、RF、SVM、XGBoost	341 例 AE	[49]
PM、HP	识别 PM 和 HP	RF	62 例 PM、60 例 HP	[50]

2 化学计量学在常见复杂疾病中的应用

常见的复杂或多因素疾病被定义为发病率较高、由多种遗传和环境因素共同决定的疾病，这类疾病通常涉及多个基因和环境影响的交互作用，如泌尿系结石、心血管疾病、糖尿病、类风湿性关节炎和精神分裂症^[51]。相对罕见病数据来说恰好相反，这类疾病数据可具备大样本量。尽管具有充足样本量和众多检测影响因素的技术和策略，但这些技术和策略具有固有的局限性，只有少数与复杂疾病有关的关键因素被确定。其发病机制通常涉及多个相互作用的因素，任何一个因素的影响都可能被其他因素掩盖或混淆，例如，泌尿系结石疾病的发生可能与饮食习惯、运动量、遗传背景等多重因素交织在一起^[52-53]。这对常见复杂疾病的筛查与诊断以及生物标志物的识别都是艰巨的任务。由于其病因复杂，往往需要采用多种研究方法和分析技术，以便全面识别潜在的生物标志物和致病机制。化学计量学等先进技术的应用，为常见复杂疾病的研究提供了新的视角和工具，有助于提升对疾病机制的理解，推动精准医疗的发展。

2.1 泌尿系结石

泌尿系结石是一种综合的慢性代谢性疾病，与心血管疾病、糖尿病等代谢综合症相关，近年来成为一个跨越所有地理区域的全球性问题，估计年发病率为 1%，患病率为 3%~5%，终生风险为 15%~25%。一旦受到影响，该病在大多数情况下往往会复发^[54-55]。泌尿系结石的发病机制尚不明确，现代医学越来越关注人体综合代谢对泌尿系结石生成的影响，为了防止或尽量减少结石的复发，许多研究将化学计量学应用于泌尿系结石疾病的分类识别、标志物鉴定以及风险评估。同时，为了满足其高维、大样本量的数据特征提取及精确分类要求，需要应用许多不同的化学计量学方法，例如传统的机器学习算法、集成学习方法和关联规则学习，选择最佳方法。

本课题组^[56]尝试使用 VIP-PLSDA、LASSO-LR 和 RFE-SVM 进行变量筛选并建立泌尿系结石风险评估模型，其中 PLSDA 表现最佳，共识别出 29 种代谢物，富集于精氨酸代谢、脂肪酸降解等关键代谢途径。通过对这些代谢物的深入分析，研究探讨了结石如何影响肾脏代谢，进而可能导致的尿液组成变化和结石形成风险。Gao 等^[57]利用一种羟基-L-脯氨酸(HLP)诱导的 CaOx 肾结石的大鼠模型，使用无监督 PCA 观察每组样本之间的分离趋势和代谢轨迹并消除异常值。为了可以在每个时间点区分 HLP 处理大鼠和对照大鼠的候选代谢物，使用 OPLS-DA 在血液和尿液中共鉴定出 55 种显著差异代谢物，并根据 R^2 、 Q^2 值和 p 值形式的交叉验证方差分析(CV-ANOVA)对 OPLS-DA 模型进行评估。Thongprayoon 等^[58]开展了一项靶向代谢组学研究(858 名受试者参与)，通过 NMR 量化的尿液代谢组学概况和根据年龄和性别调整的尿液化学成分开发模型，应用具有嵌套交叉验证的梯度提升机(GBM)方法预测结石形成状态，其平均 AUC 为 0.78。除了靶向代谢组学，PLS-DA 和 OPLS-DA 在非靶向代谢组学中也得到广泛应用，靶向代谢组学侧重于一组已知的代谢物，而非靶向代谢组学允许更全面地评估代谢组学特征^[59]。目前的靶向研究已经允许鉴定数百种化合物，然而，在非靶向研究中分析的化合物数量远远大于靶向研究的化合物数量。一般情况下，需要处理数千个代谢物信号在内的整个数据集，而其中很少有最终被确定为候选生物标志物。因此，非靶向研究需要对数据进行高度广泛的处理，PLS-DA 针对非靶向研究也能准确得到重要代谢物(疾病标志物)的重要性。例如通过兔膀胱结石模型，研究人员发现结石出现溶解现象，从而利用 PLS-DA 进行非靶向研究，筛选出 5 种特征变量，将这些差异表达的代谢物在大鼠模型中进行体内实验，证实 PLS-DA 筛选出的代谢标志物具有解释疾病病理机制的特异性和稳健性^[60]。

针对常规分类的方法，部分研究还开发出更加稳健准确的新方法，数据挖掘技术有助于疾病的早期预测并降低其发病率和相关成本。Kazemi 等^[61]基于收集的 936 名肾结石患者的特征信息，使用贝叶斯模型、SVM、不同类型的 DT、人工神经网络(ANN)和基于规则的分类器等各种数据挖掘算法进行分析，同时，该研究还提出了 4 种基于集成学习的模型，以提高每种学习算法的准确性。此外，还提出了一种在集成学习中组合单个分类器的新技术，这种技术根据遗传算法(GA)的方法给每个单独的分类器分配权重。经过 10 折交叉验证和准确性评估测试，证实该技术是一个稳健的模型(准确率为 97.1%)，可以稳定地预测患肾结石的概率。Gulhane 等^[62]在实验中将 RF 用于特征选择，确定整个训练阶段的特征相关性，这使系统能够自动选择准确预测所需的因素并确定其优先级。堆叠策略作为一种有效的集

成学习方法,能够提高分类模型的预测准确性。堆叠可合并多种分类器(LR、SVM、RF、梯度提升分类器(GBC)和多层感知器(MLP)分类器),通过添加两个额外的卷积层和两个额外的池化层来增强深度神经网络,训练过程还涉及将优化技术从随机梯度下降(SGD)更改为ADAMW(Adam with Weight Decay)。ADAMW是Adam(Adaptive moment estimation,自适应矩估计)优化器的增强版本,它将权重衰减直接整合到优化过程中,增强网络捕获输入数据中详细模式和特征的能力,这对于肾结石检测等复杂任务尤其有利。改进的DNN表现出卓越的性能,准确率为90%,精度为89%,召回率为90%,F1评分为89.5%。通过结合传统的机器学习和深度神经网络,这种集成方法可以识别数据集中的复杂模式。通过减少偏差和变异来提高疾病预测的准确性,集成方法已被证明是提高预测准确性或将复杂、困难的学习问题分解为更容易的子问题的有效方法,也是识别疾病的最佳方法^[63-64]。

2.2 糖尿病

糖尿病是一种常见的慢性疾病,属于代谢紊乱,其特征是慢性高血糖症(血糖水平高于正常人)。这是由胰岛素分泌缺陷或其生物效应受损引起的,可导致微血管(视网膜病变、肾病、神经病变)和大血管(缺血性心脏病、中风、外周血管疾病)并发症^[65]。糖尿病致使各种组织的慢性损伤和功能障碍,尤其是眼睛、肾脏、心脏、血管和神经。糖尿病可分为两类,1型糖尿病(T1DM)和2型糖尿病(T2DM),二者都受到复杂因素的相互影响^[66]。早期诊断对这两种疾病都至关重要,许多并发症可以通过早期干预来避免或减少^[67]。

Furman等^[68]运用多变量分析方法对FTIR-MIR进行了研究。通过PCA实现了对样本的聚类,有效地根据糖尿病和血脂异常的状况对样本进行分组。此外,PLS-DA模型在糖尿病诊断方面展现了超过95%的准确率,在血脂异常类型的诊断上也达到了90%以上的准确率。Al-Akl等^[69]通过PCA检测到数据中存在两个异常值并将其排除在进一步分析之外。OPLS-DA成功识别T2DM患者是否患有认知能力下降,并鉴定出20个差异表达代谢物,其中7个上调13个下调,有望成为糖尿病相关认知能力下降患者的诊断生物标志物。标志物对糖尿病的诊断和预后至关重要,合适的特征选择则可帮助识别和保留这些重要标志物,从而提高模型性能并降低复杂度。Bagherzadeh-Khiabani等^[70]基于一个由803名具有55个特征的糖尿病前期女性组成的临床数据集,通过比较几种常见的特征选择算法(包装器和过滤方法)来预测糖尿病。结果显示,包装器方法达到了最佳性能。此外,在使用的过滤方法中,对称不确定性实现了最佳的预测精度。此外,Boruta包装器算法也被用于糖尿病的特征选择,在选出所有与响应变量相关的特征后,使用线性核支持向量机(SVM-linear)、径向基核支持向量机(SVM-RBF)、k-NN、ANN和MDR算法来检测糖尿病。结果显示,SVM-linear和k-NN具有最高的AUC,被认为是该数据集的最优分类器。

为了克服单一的机器学习模型存在的局限性(如过拟合等),有研究人员结合多种算法各自的优势,获得了更高的预测准确性。Ahmed等^[71]提出了一种糖尿病预测融合模型(FMDP),使用SVM和ANN模型进行预测,在模型融合阶段,通过模糊规则(Fuzzy rules)集成两种机器学习的输出进行最终预测。所提出的模糊决策系统达到了94.87%的准确率,高于其他现有系统。Polat等^[72]在广义判别分析(GDA)和最小二乘支持向量机(LS-SVM)的基础上,提出了一种新型的级联学习系统。首先,使用GDA预处理初始数据并区分健康和糖尿病患者;随后,使用LS-SVM对糖尿病数据集进行分类。结果显示分类准确率为82.05%,与其他分类技术报道的先前结果相比,具有很好的应用前景。与其他机器学习模型相比,集成模型使用先进的提升技术进行正则化,限制过拟合问题并得到更高的准确性。

近年来,随着科技的发展,各类组学数据(如基因组学、蛋白质组学、转录组学和代谢组学)的获取变得越来越容易,可用性显著提高。多组学研究通过结合不同组学的数据,展示了巨大的潜力,能够深入揭示健康与疾病状态下的分子机制,并绘制生物体的相互作用图谱,从而为糖尿病的理解和管理提供全面的视角^[73]。Carrasco-Zanini等^[74]通过有效整合多种生物信息,使用最小绝对收缩和选择算子(LASSO)回归从5759个基因组、蛋白质组、代谢组和临床生物标志物特征中选出最重要的预测因子,建立了预测T2DM发生模型,并与现有临床模型的预测性能进行对比,发现稀疏模型可降低大型组学特征中验证和协调检测带来的成本和需求,提高更具包容性的特征被转化为临床环境的可能性。

进一步地,随着深度学习技术在医疗领域的广泛应用,该类方法也开始应用于糖尿病诊断和预测。

Dudukcu 等^[75]使用长短期记忆(LSTM)、WaveNet 和门控循环单元(GRU)以及这些架构的决策级组合来预测血糖，为这些网络架构的最有效运行选择超参数，并使用扩展的 OhioT1DM 数据集进行实验研究，通过 3 个网络融合获得平均最低误差值。与其他文献研究相比，所提出的融合方法为血糖预测的未来研究提供了新的视角，能够预防未来血糖发作的潜在危险；在预测准确性提高方面，模型训练和预测过程中系统带来的计算成本可以忽略不计。Alhassan 等^[76]介绍了糖尿病研究中最大的患者记录数据集 KAIMRCD，该数据集包含约 14 609 例病例的全面信息。为了准确诊断 T2DM，将数据集作为时间序列进行处理，使用时间预测深度学习模型进行研究，提出的 LSTM 和 GRU 模型仅需 3 个时间序列且无需复杂特征工程，即可实现极高准确率(超过 97%)。Nilashi 等^[77]开发了一种综合了奇异值分解(SVD)、SOM、逐步判别分析(STEPDISC)和深度信念网络(DBN)的新集成技术进行糖尿病诊断。采用 SVD 进行降维和缺失值预测，SOM 对数据进行聚类，STEPDISC 用于特征选择，最后使用 DBN 从大量数据中学习并预测糖尿病。与文献中的其他方法相比，新集成技术呈现出更好的训练误差，并且在所有集群中都实现了快速收敛，该研究的结果可帮助医生做出正确的诊断。化学计量学在常见复杂疾病中的应用见表 2。

表 2 化学计量学在常见复杂疾病中的应用
Table 2 Applications of chemometrics in common complex diseases

Diseases	Objective	Method	Sample size	Reference
泌尿系 结石	识别健康对照和泌尿系结石患者，以全面地了解代谢扰动在生物通路中的参与	LR、SVM、PLS-DA	124 例健康对照和 85 例结石患者	[56]
CaOx 肾 结石	建立羟基-L-脯氨酸(HLP)诱导的 CaOx 肾结石大鼠模型，发现潜在的代谢标志物和代谢机制	PCA、OPLS-DA	31 只大鼠	[57]
肾结石	探索肾结石形成的尿液代谢物与化学成分之间的差异和确切机制	GBM	418 例肾结石、440 例对照	[58]
肾结石	预测肾结石的类型和影响患病风险因素	贝叶斯模型、DT、ANN 等	936 例肾结石	[61]
肾结石	通过文本数据全面了解肾脏健康状况并实现及时干预	GBM、Light GBM、CatBoost、SVC、LR、XGBoost、DNN 等	79 例肾结石	[62]
糖尿病	识别糖尿病和血脂异常患者，发现生物标志物，通过早期检测措施提高患者预后	PCA、PLS-DA	40 例非糖尿病、10 例糖尿病前期、10 例糖尿病	[68]
T2DM	确定 T2DM 相关认知能力下降(DACD)中改变的代谢途径	PCA、OPLS-DA	54 例 T2DM 认知功能正常、46 例 T2DM 认知功能下降	[69]
糖尿病	使用变量选择方法提高临床预测模型的性能	LR	803 例糖尿病前期	[70]
糖尿病	开发基于 FMDP 算法进行糖尿病预测的模型	FMDP	315 例阳性、205 例阴性	[71]
糖尿病	开发新型的级联学习系统诊断糖尿病	GDA、LS-SVM	268 例阳性、500 例阴性	[72]
T2DM	测试基因组、蛋白质组、代谢组和临床生物标志物在 T2DM 临床预测模型中的性能	LASSO 回归	1 105 例 T2DM	[74]
糖尿病	尝试使用糖尿病患者病史的血糖值来预测未来血糖值，定期监测循环系统中的血糖	LSTM、WaveNet、GRU	12 例糖尿病患者	[75]
T2DM	建立最大的糖尿病患者记录数据集，诊断 T2DM	LSTM、GRU	14 609 例 T2DM	[76]
糖尿病	进行糖尿病风险预测	SVD、SOM、PCA、STEP-DISC、DBN 等集成技术	268 例阳性、500 例阴性	[77]

3 两类疾病化学计量学方法的比较与讨论

近年来生物医学领域快速发展，化学计量学也随之发展成一门广泛、综合且非常有益的学科，在早期筛查诊断、疾病治疗靶点、复发预测、优化预后等方面发挥重要作用，包括了从单组学独立分析到多组学联合分析，从小样本个体数据和临床表型到大规模健康数据库和涉及大量患者队列的高维度研究。本研究探讨了化学计量学在罕见病与常见复杂疾病中的应用及其潜在影响。结果显示，尽管这两类疾病在样本量、数据复杂性及研究目标上存在显著差异，但化学计量学作为一种强大的数据分析工具，在两者的研究中均发挥了重要作用。

针对罕见病，小样本可导致系统和随机误差(偏差和方差)，模型稳定性(方差)既可通过比较基于略微变化的训练数据构建的多个模型的模型参数来衡量，也可通过比较这些模型对同一样本的预测结果方差进行评估。迭代k折交叉验证或 Bootstrap 等重采样技术可生成所需模型^[78]。Bootstrap 是解决罕见疾病小样本最常用的集成学习策略，可以避开众多统计分布假设，通过有放回的抽样方式生成多个

不同的训练集和测试集,广泛应用于集成学习,以提高模型的稳健性。其次,化学计量学方法能够在样本量有限的情况下提取有用信息,识别潜在的代谢标志物。这一特点在许多罕见代谢病的研究中得到了充分体现。例如,通过PCA和PLS等技术,研究者能够从有限的患者样本中发现与疾病相关的代谢物变化,从而为早期诊断和个性化治疗提供支持。然而,罕见病研究中的主要挑战在于样本的异质性和标签噪声,这可能导致结果的不确定性,使得对罕见病进行分类和预测具有挑战性^[79]。因此,未来的研究需要更加关注标准化的分析方法和严格的统计设计,以提高结果的可靠性,同时进一步促进多中心开展或采用新的机器学习和深度学习(meta-learning^[80]等)。此外,多组学融合技术通过组学之间的交叉互补来探索不同组学之间的关联性,能够弥补单一组学的信息不足,进而可为研究罕见病的内在疾病机理提供新的思路。

相对而言,常见复杂疾病的研究则呈现出另一种特征。由于样本量相对较大,使用化学计量学方法能够有效处理复杂的多变量数据。在这种情况下,除了经典的化学计量学方法外,机器学习技术如随机森林和支持向量机被广泛应用,通过成功学习数据中的复杂关联,以自动化方式分析高维和复杂数据,使研究者能够从大量待测物中筛选出与疾病相关的特征。这种方法不仅提高了对疾病机制的理解,还为疾病的早期预警和风险评估提供了重要依据。由于疾病的病因复杂以及在遗传、基因组和蛋白质组学水平上复杂的分子机制,常见复杂疾病的研究同样面临挑战,特别是不同个体之间的代谢差异和环境因素的影响,可能导致分析结果的复杂性与不确定性。同时,这些模型框架在提供建模灵活性和鲁棒性的同时,由于其复杂的算法导致存在不透明和难以解释每个单独特征的困扰^[81]。开发侧重于模型可解释性的方法可以识别可能与潜在疾病机制相关的特征,而识别重要的生物标志物是协助研究人员建立有关预防、诊断和治疗复杂人类疾病的关键性任务,可帮助了解特征如何影响模型的性能,从而深入了解其对潜在疾病机制的潜在贡献^[82]。目前已经开发了许多基于集成学习的模型,并广泛用于减轻复杂的人类疾病研究中的一些分析挑战^[81]。

化学计量学在整合多种组学数据方面的潜力同样不可忽视。结合基因组学、转录组学与代谢组学的数据分析,能够为罕见病和常见复杂疾病提供更全面的生物学视角。这种数据整合不仅有助于理解疾病的多因素背景,还可能揭示新的生物标志物和治疗靶点,为中下游的临床应用奠定基础。

值得注意的是,虽然化学计量学在临床医学中的应用前景广阔,但实际转化仍面临多重挑战。数据质量与标准化问题是首要障碍,医疗数据普遍存在噪声、缺失和格式不统一现象,尤其在影像和电子健康记录整合中表现突出。这一问题可考虑通过数据清洗、多中心标准化加以改善,但同时需平衡数据共享与隐私保护之间的矛盾。技术层面,由于医疗决策的高风险特性,医生对算法模型的准确性普遍持审慎态度,需结合专家知识优化模型的可解释性,并通过严格质量控制和专家复核机制降低误诊可能,增强临床医生对模型诊断结果的信任。此外,应用场景针对性较强的模型泛化能力不足限制了其跨机构应用,而迁移学习和多阶段验证或可提升适应性。

总体而言,尽管大量证据表明,化学计量学方法在疾病的标志物挖掘、病理机制探究以及辅助诊断等方面发挥了重要作用,但真正落实进入临床环境仍然需要更多的支持,以弥合技术与医疗现实的差距,这需要多方持续努力,使该方法更接近真正临床应用的目标。

4 结 论

化学计量学在罕见疾病及常见复杂疾病的研究中发挥了重要作用,成为理解疾病机制和促进个性化医疗的关键工具。通过有效整合和分析高维组学数据,化学计量学不仅揭示了潜在的生物标志物,还推动了早期诊断和靶向治疗的发展。机器学习的引入进一步提升了数据分析的能力,使研究者能够从复杂的实验数据中提取有用的特征和规律。尽管目前仍面临一些挑战,如数据的异质性和模型的可解释性,但随着技术的不断进步,化学计量学有潜力为疾病的早期诊断、个性化治疗及公共健康策略的制定提供重要支持。未来的研究应继续探索优化算法与模型,以提高对疾病机制的理解和临床应用的转化效率。

参考文献:

- [1] Kowalski B R. *J. Chem. Inf. Comput. Sci.*, **1975**, 15(4): 201-203.

- [2] Stanimirova I, Daszykowski M, Hopke P K. *TrAC, Trends Anal. Chem.*, **2024**, 173: 117642.
- [3] Paul A, Harrington P B. *TrAC, Trends Anal. Chem.*, **2021**, 135: 116165.
- [4] Wold S. *Chemom. Intell. Lab. Syst.*, **1995**, 30(1): 109–115.
- [5] Xiong X, Zhu Q, Zhou Z, Qian X, Hong R, Dai Y, Hu C. *Clin. Radiol.*, **2023**, 78(11): e839–e846.
- [6] Dumas M E. *Mol. Biosyst.*, **2012**, 8(10): 2494–2502.
- [7] Frédérick M, Pirotte B, Fillet M, de Tullio P. *J. Med. Chem.*, **2016**, 59(19): 8649–8666.
- [8] Tebani A, Abily–Donval L, Afonso C, Marret S, Bekri S. *Int. J. Mol. Sci.*, **2016**, 17(7): 1167.
- [9] Bracewell–Milnes T, Saso S, Abdalla H, Nikolau D, Norman–Taylor J, Johnson M, Holmes E, Thum M Y. *Hum. Reprod. Update*, **2017**, 23(6): 723–736.
- [10] Puthongkham P, Wirojsaengthong S, Suea–Ngam A. *Analyst*, **2021**, 146(21): 6351–6364.
- [11] Fisher R A. *Ann. Eugen.*, **1936**, 7(2): 179–188.
- [12] Boulesteix A L. *Stat. Appl. Genet. Mol. Biol.*, **2004**, 3(1): 1–30.
- [13] Zhang L, Zhou W D. *Inform. Sci.*, **2016**, 343/344: 79–93.
- [14] Booker N K, Knights P, Gates J D, Clegg R E. *Eng. Fail. Anal.*, **2022**, 132: 105937.
- [15] Schieppati A, Henter J I, Daina E, Aperia A. *Lancet (London, England)*, **2008**, 371(9629): 2039–2041.
- [16] Schaefer J, Lehne M, Schepers J, Prasser F, Thun S. *Orphanet J. Rare Dis.*, **2020**, 15(1): 145.
- [17] Ni X, Shi T. *Sci. China Life Sci.*, **2017**, 60(7): 681–685.
- [18] Brereton R G, Jansen J, Lopes J, Marini F, Pomerantsev A, Rodionova O, Roger J M, Walczak B, Tauler R. *Anal. Bioanal. Chem.*, **2018**, 410(26): 6691–6704.
- [19] Mak C M, Lee H C, Chan A Y, Lam C W. *Crit. Rev. Clin. Lab. Sci.*, **2013**, 50(6): 142–162.
- [20] Blekherman G, Laubenbacher R, Cortes D F, Mendes P, Torti F M, Akman S, Torti S V, Shulaev V. *Metabolomics*, **2011**, 7(3): 329–343.
- [21] Moreda–Piñeiro J, Cocho J A, Couce M L, Moreda–Pineiro A, Bermejo–Barrera P. *Metallomics*, **2021**, 13(5): mfab020.
- [22] Zaunseder E, Mutze U, Garbade S F, Haupt S, Feyh P, Hoffmann G F, Heuveline V, Kolker S. *Metabolites*, **2023**, 13(2): 304.
- [23] Liu H, Zhu J, Li Q, Wang D, Wan K, Yuan Z, Zhang J, Zou L, He X, Miao J. *Funct. Integr. Genomics*, **2021**, 21(5/6): 645–653.
- [24] Kalivodová A, Hron K, Filzmoser P, Najdekr L, Janečková H, Adam T. *J. Chemom.*, **2015**, 29(1): 21–28.
- [25] Wang B, Liu G, Fei Q, Zuo Y, Ren Y. *Spectrochim. Acta A*, **2009**, 71(5): 1695–1700.
- [26] Weerd J C V, Wegberg A, Boer T S, Engelke U F H, Coene K L M, Wevers R A, Bakker S J L, Blaauw P, Groen J, Spronsen F J V, Heiner–Fokkema M R. *Metabolites*, **2024**, 14(9): 479.
- [27] Yang Q, Lin S S, Yang J T, Tang L J, Yu R Q. *Talanta*, **2017**, 165: 545–552.
- [28] Chen N, Chen S, Zhang Q, Wang S R, Tang L J, Jiang J H, Yu R Q, Zhou Y P. *Microchem. J.*, **2023**, 190: 108600.
- [29] Yang Q, Tan L, Wu B Q, Tian G L, Xu L, Yang J T, Jiang J H, Yu R Q. *Chemom. Intell. Lab. Syst.*, **2020**, 204: 104107.
- [30] Zhu Z X, Genchev G Z, Wang Y M, Ji W, Ren Y Y, Tian G L, Sriswasdi S, Lu H. *World J. Pediatr.*, **2024**, 20(10): 1090–1101.
- [31] Yang Q, Tian G L, Qin J W, Wu B Q, Tan L, Xu L, Wu S Z, Yang J T, Jiang J H, Yu R Q. *Talanta*, **2020**, 219: 121370.
- [32] Yang Q, Shi B H, Tian G L, Niu Q Q, Tang J, Linghu D D, He H Q, Wu B Q, Yang J T, Xu L, Yu R Q. *Microchem. J.*, **2021**, 168: 106350.
- [33] Xiao W, Niu Q Q, Sun Z Y, Xiong J, Yang Q, Wu B Q, Xu L, Yang Y L, Yu R Q. *Microchem. J.*, **2022**, 181: 107754.
- [34] Trygg J. *J. Chemom.*, **2002**, 16(6): 283–293.
- [35] Löfstedt T, Trygg J. *J. Chemom.*, **2011**, 25(8): 441–455.
- [36] Löfstedt T, Hoffman D, Trygg J. *Anal. Chim. Acta*, **2013**, 791: 13–24.
- [37] Singh A, Shannon C P, Gautier B, Rohart F, Vacher M, Tebbutt S J, Le Cao K A. *Bioinformatics*, **2019**, 35(17): 3055–3062.
- [38] Li Z Y, Liu X Y, Xiao W, Yang J T, Jiang P P, Wu B Q, Liu X J, Xue M, Lv H J, Zhou S H, Yang Q, Xu L, Yang Y L. *Microchem. J.*, **2024**, 207: 112018.
- [39] Usha Rani G, Kadali S, Kurma Reddy B, Shaheena D, Naushad S M. *Metabolomics*, **2023**, 19(5): 49.
- [40] Leal–Witt M J, Rojas–Agurto E, Munoz–Gonzalez M, Penaloza F, Arias C, Fuenzalida K, Bunout D, Cornejo V, Acevedo A. *Metabolites*, **2023**, 13(6): 677.
- [41] Li H, Gao S, Wu D, Zhu M, Hu Z, Fang K, Chen X, Ni Z, Li J, Zhao B, She X, Huang X. *Int. J. Med. Inf.*, **2025**, 195: 105765.
- [42] Groen J, de Haan B M, Overduin R J, Haijer–Schreuder A B, Derks T G J, Heiner–Fokkema M R. *Orphanet J. Rare Dis.*, **2025**, 20(1): 15.

- [43] Messa G M, Napolitano F, Elsea S H, di Bernardo D, Gao X. *Bioinformatics*, **2020**, 36(S2): i787–i794.
- [44] Davidson A, Diamond B. *N. Engl. J. Med.*, **2001**, 345(5): 340–350.
- [45] Cooper G S, Stroehla B C. *Autoimmun. Rev.*, **2003**, 2(3): 119–125.
- [46] Fugger L, Jensen L T, Rossjohn J. *Cell*, **2020**, 181(1): 63–80.
- [47] Xiong W, Yeo T, May J T M, Demmers T, Ceronie B, Ramesh A, McGinty R N, Michael S, Torzillo E, Sen A, Anthony D C, Irani S R, Probert F. *Ann. Clin. Transl. Neurol.*, **2024**, 11(7): 1897–1908.
- [48] Guo J, Dong R, Zhang R, Yang F, Wang Y, Miao W. *J. Affect. Disord.*, **2025**, 369: 352–363.
- [49] Zhang Y, Shi X, Fan Z, Tu E, Wu D, Leng X, Wan T, Wang X, Wang X, Lu W, Du F, Jiang W. *J. Psychosom. Res.*, **2025**, 190: 112051.
- [50] Mekki A, Dercle L, Lichtenstein P, Nasser G, Marabelle A, Champiat S, Chouzenoux E, Balleyguier C, Ammari S. *Eur. J. Cancer*, **2019**, 119: 44–56.
- [51] Thomson G, Esposito M S. *Trends Genet.*, **1999**, 15(12): M17–M20.
- [52] Schork N J. *Am. J. Respir. Crit. Care Med.*, **1997**, 156(4): S103–S109.
- [53] Peerapen P, Thongboonkerd V. *Adv. Nutr.*, **2023**, 14(3): 555–569.
- [54] Moe O W, Pearle M S, Sakhaee K. *Kidney Int.*, **2011**, 79(4): 385–392.
- [55] Scales C D Jr., Tasian G E, Schwaderer A L, Goldfarb D S, Star R A, Kirkali Z. *Clin. J. Am. Soc. Nephrol.*, **2016**, 11(7): 1305–1312.
- [56] Yang J, Zhang D, Lu Y, Mai H, Wu S, Yang Q, Zheng H, Yu R, Luo H, Jiang P, Wu L, Zhong C, Zheng C, Yang Y, Cui J, Lei Q, He Z. *Kidney Blood Press. Res.*, **2024**, 50(1): 83–96.
- [57] Gao S, Yang R, Peng Z, Lu H, Li N, Ding J, Cui X, Chen W, Dong X. *Sci. Rep.*, **2016**, 6: 30142.
- [58] Thongprayoon C, Vuckovic I, Vaughan L E, Macura S, Larson N B, D'Costa M R, Lieske J C, Rule A D, Denic A. *J. Am. Soc. Nephrol.*, **2022**, 33(11): 2071–2086.
- [59] Gorrochategui E, Jaumot J, Lacorte S, Tauler R. *TrAC, Trends Anal. Chem.*, **2016**, 82: 425–442.
- [60] Zhang X Z, Lei X X, Jiang Y L, Zhao L M, Zou C Y, Bai Y J, Li Y X, Wang R, Li Q J, Chen Q Z, Fan M H, Song Y T, Zhang W Q, Zhang Y, Li L J, Xie H Q. *Signal Transduction Targeted Ther.*, **2023**, 8(1): 41.
- [61] Kazemi Y, Mirroshandel S A. *Artif. Intell. Med.*, **2018**, 84: 117–126.
- [62] Gulhane M, Kumar S, Choudhary S, Rakesh N, Zhu Y, Kaur M, Tandon C, Gadekallu T R. *SLAS Technol.*, **2024**, 29(4): 100159.
- [63] Mahajan P, Uddin S, Hajati F, Moni M A. *Healthcare*, **2023**, 11(12): 1808.
- [64] Krawczyk B, Minku L L, Gama J, Stefanowski J, Woźniak M. *Inform. Fusion*, **2017**, 37: 132–156.
- [65] Cobelli C. *J. Phys. Conf. Ser.*, **2010**, 238(1): 012003.
- [66] Forouhi N G, Wareham N J. *Medicine (Baltimore)*, **2014**, 42(12): 698–702.
- [67] Madsen R, Lundstedt T, Trygg J. *Anal. Chim. Acta*, **2010**, 659(1/2): 23–33.
- [68] Furman A E F, Cobre A D F, Stremel D P, Pontarolo R. *Chemom. Intell. Lab. Syst.*, **2024**, 252: 105179.
- [69] Al-Akl N S, Khalifa O, Ponirakis G, Parray A, Ramadan M, Khan S, Chandran M, Ayadathil R, Elsotouhy A, Own A, Al Hamad H, Decock J, Alajez N M, Albagha O, Malik R A, El-Agnaf O M A, Arredouani A. *Int. J. Mol. Sci.*, **2024**, 25(4): 2247.
- [70] Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg E W, Khalili D. *J. Clin. Epidemiol.*, **2016**, 71: 76–85.
- [71] Ahmed U, Issa G F, Khan M A, Aftab S, Khan M F, Said R A T, Ghazal T M, Ahmad M. *IEEE Access*, **2022**, 10: 8529–8538.
- [72] Polat K, Güneş S, Arslan A. *Expert Syst. Appl.*, **2008**, 34(1): 482–487.
- [73] Smith G D, Ebrahim S. *Int. J. Epidemiol.*, **2003**, 32(1): 1–22.
- [74] Carrasco-Zanini J, Pietzner M, Wheeler E, Kerrison N D, Langenberg C, Wareham N J. *Diabetologia*, **2024**, 67(1): 102–112.
- [75] Dudukcu H V, Taskiran M, Yildirim T. *Biocybern. Biomed. Eng.*, **2021**, 41(3): 1208–1223.
- [76] Alhassan Z, McGough A S, Alshammari R, Daghestani T, Budgen D, Al Moubayed N. *Artificial Neural Networks and Machine Learning–ICANN 2018*, **2018**, 11141: 468–478.
- [77] Nilashi M, Abumalloh R A, Alyami S, Alghamdi A, Alrizq M. *Diagnostics*, **2023**, 13(10): 1821.
- [78] Beleites C, Salzer R. *Anal. Bioanal. Chem.*, **2008**, 390(5): 1261–1271.
- [79] Banerjee J, Taroni J N, Allaway R J, Prasad D V, Guinney J, Greene C. *Nat. Meth.*, **2023**, 20(6): 803–814.
- [80] Singh K, Malhotra D. *Arch. Comput. Methods Eng.*, **2023**, 30(7): 4081–4112.
- [81] Mi X, Zou B, Zou F, Hu J. *Nat. Commun.*, **2021**, 12(1): 3008.
- [82] Lundberg S M, Nair B, Vavilala M S, Horibe M, Eisses M J, Adams T, Liston D E, Low D K W, Newman S F, Kim J, Lee S I. *Nat. Biomed. Eng.*, **2018**, 2(10): 749–760.