DOI：10.12287/j.issn.2096-8965.20220309

·守正创新·

中医表型组学的数据特征与研究方法展望

左玲，陈建新*

（北京中医药大学中医学院，北京 100029）

【摘要】人类表型组学构建了跨尺度组学数据的关联，对精准医学的实现具有重要价值。中医学与表型组学都关注人体复杂系统中各种表型的变化规律，二者的理念与技术可以相互促进和补充，中医表型组学研究将有助于实现中西医融合协同发展和促进。本研究首先总结了现有中医特色表型的数据特征、表型收集和量化的方法，随后从大数据和人工智能角度简述中医表型的数据处理、跨尺度关联和精细调控机制的手段，最后提出了现存的挑战与未来的研究方向，以期为中医表型组学研究者提供参考。

【关键词】中医表型组学；表型组学；人工智能；中医药

中图分类号：R2-03　　**文献标识码**：A　　**文章编号**：2096-8965（2022）03-0064-05

Traditional Chinese medicine phenomics: data characteristics and prospect of research methods

Zuo Ling, Chen Jianxin*

(School of Traditional Chinese Medicine, Beijing University of Traditional Chinese Medicine, Beijing 100029, China)

【**Abstract**】Human phenomics constructs the association of cross-scale-omics data, which is of great value to the realization of precision medicine. Both TCM and phenomics focus on the variation patterns of various phenotypes in the complex system of human body, and the concepts and techniques of both can promote and complement each other, and the study of TCM phenomics will help to realize the synergistic development and promotion of the integration of Chinese and Western medicine. This study first summarizes the existing data characteristics of TCM characteristic phenotypes, methods of phenotype collection and quantification, and then briefly describes the means of data processing, cross-scale correlation and fine regulation mechanisms of TCM phenotypes from the perspective of big data and artificial intelligence, and finally proposes the existing challenges and future research directions in order to provide reference for TCM phenomics researchers.

【**Keywords**】Traditional Chinese medicine phenomics; Phenomics; Artificial intelligence; Traditional Chinese medicine

人类表型组学是当前生命科学领域的新兴学科，通过全景表型测量与数据整合分析，旨在构建基因组、蛋白组、代谢组等微观表型数据到遗传性状、生理状态、行为功能等宏观表型数据的跨尺度关联[1]。随着大数据和人工智能科学技术不断发展，寻找跨尺度的海量数据之间的关联成为可能。自人类表型组学概念提出以来，已搭建起一套精密规范的测量平台，并逐渐形成一系列人类表型图谱。人类表型组学可以深入研究多种组学数据对健康和疾病的影响和关联，被国际科学家认为是实现精准医

收稿日期：2022-06-14；**修回日期**：2022-07-11

基金项目：国家自然科学基金项目（82025036）；国家重点研发计划（2017YFC1700106）

通讯作者：陈建新（1981-），男，福建莆田人，研究员，北京中医药大学二级教授，主要从事中医信息学研究。E-mail：cjx@bucm.edu.cn

学的七大机会之一[2]，对精准治疗疾病和健康管理有着重要意义。表型组学研究需要从大量生物表型中提取线索，而中医学通过千百年发展，积累了大量临床资料和对表型的独特的认识。多组学数据和中医表型的交叉碰撞有利于中西医融合发展。但当前许多中医研究虽然运用了组学技术，但缺乏宏观表型和微观表型之间的跨尺度联系，导致研究成果难以回归临床[3]。因此，研究中医学中各种跨尺度表型关联与机制的中医表型组学逐渐兴起[4]。当前，中医表型组学尚处于初期建立阶段，其主要特点是以中医理论为核心，以中医学病证发展过程中的表型为研究对象，基于大样本队列和多组学跨学科技术等手段[3, 5, 6]，通过研究中医表型的跨尺度关联和机制，阐释中医理论内涵。本研究旨在总结中医现有特色表型的数据现状特征，并从大数据和人工智能角度简述中医表型数据处理的方法与现存问题，以期为中医表型组学研究者提供参考。

1 中医表型的数据现状及特征

传统中医学通过“望闻问切”四诊采集病人的外部资料，通过中医理论分析归纳并不断总结，最终提炼出一套特殊的表型集合来反映疾病的外在表现和内在本质。从数据结构特点，我们认为中医表型分为简单表型和复杂表型两类，前者指可以分为面色、舌色、舌齿痕、掌纹、脉象等单一表型；而复杂表型是将各种单一表型重组整合得到的一系列表型集合，如证候、体质等[5]。当前，四诊简单表型的客观化仍是中医学领域的重点和难点研究方向，而复合表型的构建和整合还缺乏规范化和标准化探索。

1.1 简单表型

当前，四诊的客观化和标准化研究已经取得了较大进展。望诊是当前四诊客观化积累最多的领域，而又尤以面象和舌象最为热门。舌象可提取的特征表型数据包括舌色、苔色、舌苔特征、舌型、舌态等[7]。而面象特征客观化则以面部色彩为主，也有少量研究应用红外成像、广电血流容积仪等技术量化面部特征，但仍未形成广泛应用[8–10]。当前望诊数据客观化存在的主要问题包括：一是数据质量参差不齐，采集设备不统一；二是量化值与中医认识不一致，影响分类准确性，如颜色空间坐标与临床认识不完全相同，认可度不高，最终输出表型往往仍是分类结果，损失了量化信息；三是特征提取技术还有待提升，如舌齿痕、裂纹等高级图像特征受到边缘分割算法、目标检测算法等技术限制，提取效果和准确度不如色彩值。

闻诊表型数据方面，主要分为声学数据和气味数据两类。其中声学表型客观化研究整体较少，主要是通过采集音频并提取波谱频段特征信号[11]，但同样存在采集仪器不统一、数据处理未形成范式的问题，离临床辅助诊断应用还有相当大的技术壁垒。而气味数据收集与量化至今仍鲜有报道，近期一篇针对人类气味表型驱动蚊子行为的研究提供了一套人类气味表型收集和量化的设计思路，可以通过空气流动和收集装置收集人体特有气味并提取其中的标志性气味分子[12]，填补了气味表型研究的空白。

脉诊表型数据的客观化研究进展则主要依赖于脉象传感器的设计和信号处理技术，现有的脉诊仪传感器可根据工作原理不同分为压力式、光电式、超声式等[13, 14]，可以模仿三部九候方式动态采集脉冲波形信号[15]，并结合机器学习和深度学习方法建立分类模型[16]。然而，受限于传感器的精度和训练集的数量，脉象信息的获取和采集精度仍然不足。此外，脉象的传统分类依赖于医生经验，脉象仪的分类结论难以得到中医研究者的广泛认同，而脉象图等非传统数据类型则处于初期积累阶段，缺乏统一标准[14, 17]。

问诊表型由于常涉及多种表型的综合采集，且形式多以问卷和量表设计为主，在表型提取形式上与复合表型更为类似。

1.2 复合表型

常见的中医复合表型包括证候、体质，以及问诊中的多种体征、症状等，对复合表型的收集方式目前主要有以下几类：

1.2.1 基于中医传统理论经验、专家共识等行业标准或国家标准的人工标注，多见于临床各种病–证治疗研究等。例如，病证结合研究中，需要在特定疾病下收集证候表型，而现有的常见做法是由临床医生参照行业内部的辨证标准或专家共识进行证候分类判别。例如血瘀证辨证时，满足舌质紫暗、瘀斑瘀点、肌肤甲错、固定性疼痛等12条临床表型中的2项即可诊断为血瘀证[18, 19]。但细究其中，一方面，辨证标准的制定过程较为模糊，对证候这种复杂表型中的简单表型并未作精确规定，也缺乏量化标准参考，因此在实际操作中存在大量主观因素影响。另一方面，中医业界对“辨证标准”的内涵存

在歧义，“证”本质的现代化研究停留于探索阶段，难以形成公认的评价体系，而不同临床经验的医生对证型的内涵理解存在差异，这导致医生对单一表型在构成复杂表型时的权重认识不一致，仍缺乏业内广泛认可的诊断标准[20]。

1.2.2 基于流行病学调查或临床资料数据和统计学方法开发的量表或问卷数据，例如证候诊断标准、中医体质问卷等。由于传统理论和专家共识的表型收集方法仍不够客观化和标准化，对于复杂的中医表型已经开发出越来越多的诊断量表或问卷。常见的量表构建方法包括主观的德尔菲法，客观的克朗巴赫系数法、因子分析法、聚类法等统计学方法，以及主客观结合法等[21, 22]。这一类最具代表性的是由王琦院士团队开发的中医体质学问卷，通过流行病学和临床调查[23-25]、统计学信效度检验和准确性评价[26, 27]，开发出了一套成熟标准的中医体质分类问卷，为中医体质表型的提取打下了坚实基础。但当前成熟可靠的量表仍较少，多数研究者并未考虑到量表中单一表型（条目）在复杂表型中的权重[21]，在最终分类的准确性上仍有提升空间。另外大多数量表在制定时并未进行信度、效度检验，更鲜有研究报告反应度和可行性等，这影响了量表的可信程度，难以进行大规模推广。

1.2.3 基于表型组学思路的跨尺度表型客观化收集和大数据整合。当前，基于表型组学研究中医问题的思路尚处于起步阶段，有巨大的研究潜力。随着中医诊断学的不断进步，研究者们认识到对于中医宏观表型，应基于足够客观、准确、稳定收集的四诊信息，才能确保表型具有较好的可重复性，从而规范化推断出复杂表型[28]。此外，当前的中医诊断量表或问卷收集的表型中绝大多数都是临床症状和体征等宏观表型，越来越多的学者认识到应该吸纳现代科学技术和方法，补充中医复杂表型的微观认识[20]。当前直接对中医复杂表型进行整合的方法仍较少，如基于熵的复杂系统划分方法制定证候量化标准等，但现代生物医学已有一些统计学建模或遗传网络构建[29-31]方法，可以客观化构建复杂的临床表型，为中医表型组学的研究提供了一定参考。值得注意的是，当前已有一系列数据库提供了常见中医症状、证型等表型与现代医学症状、疾病关联的数据库，如北京中医药大学开发的Symmap数据库[32]、中国中医科学院中药研究所开发的SoFDA平台[33]等，为挖掘病-证关系、挖掘中医表型的跨尺度表型提供了技术和数据支持。

2 中医表型组学的数据处理及分析方法

2.1 中医表型的数据处理与人工智能算法

中医表型数据的量化是当前中医表型组学发展的重要问题之一。以舌象为例，舌诊是当前中医表型客观化研究的热点方向，其数据采集流程包括：（1）数据采集：使用相机、舌诊仪等图像采集仪器收集舌头图像；（2）数据预处理和质控：对于非标准设备采集的图像，需要对环境光线、色差、对比度等进行校准；（3）数据特征提取，如舌色、苔色、舌形、舌齿痕、舌裂纹、舌点刺等。

数据特征提取是中医表型量化的难点问题。对于低级图像信息如颜色，可以引入色值空间、色彩空间等成熟的提取器进行检测。但对于更高级的图像特征，需要借助人工智能算法如机器学习或深度学习来获取。传统的机器学习算法包括决策树、随机森林、SVM、k-NN、贝叶斯算法、k-means聚类等[34, 35]。通过对设计特征提取器检测目标表型，如舌形可以借助边缘分割算法获取[36]，舌部纹理可以通过局部特征的灰阶差分参数[37]提取识别，舌色分类则可以借助RF算法[34]。机器学习模型的优点在于并不需要特别大的数据样本即可提取特征，但缺点在于特征提取器依赖人工设计，对特征信息细节的纳入范围过低、超参数的盲目调整、特定型号的采集设备都可能是传统机器学习过拟合的原因。受限于上述原因，传统机器学习算法的鲁棒性受到限制，跨任务迁移能力也欠佳。

随着深度学习技术的发展，以卷积神经网络（Convolutional Neural Network, CNN）为代表的深度学习模型逐渐取代传统机器学习模型被运用于中医表型的特征提取。深度学习算法作为多层学习算法，可以从原始数据中通过自适应学习逐层建立特征提取层参数，而无需人工设计特征提取器，从而提升了模型的鲁棒性和任务迁移能力。本团队之前提出的基于CNN的舌腻苔分类模型可以达到88%以上的准确度[38]，而对舌齿痕的分类模型则可以达到90%以上的准确度[39]。深度学习模型的优点是鲁棒性和迁移能力明显提升，但缺点在于对数据规模和算力要求更高、建模过程为黑箱式、参数众多且常常难以解释以及调参更为复杂等。

无论采用机器学习或深度学习，当前技术和数据仍只能做到有监督学习，而这依赖于人工建立先

验标签信息，且标注者的临床水平、知识范围将会极大影响标签的稳定性，进而影响特征提取能力和迁移能力，因此，实际操作中还需要注意对标注标准的统一和标注者的培训，以尽量提高标签的稳定性和一致性。

2.2 中医表型的跨尺度关联及精细调控机制分析

当前，以全基因组关联分析（Genome-Wide Association Studies，GWAS）、全表观遗传组关联研究（Epigenome-Wide Association Studies，EWAS）、全表型组关联分析（phenome-Wide Association Studies, pheWAS）等为代表的跨尺度组学关联分析已经取得大量数据积累并逐渐形成了稳定的分析范式。这些分析方法桥接了表型与基因组学、表观基因组等微观组学数据，从而筛选与宏观表型显著相关的遗传物质或分子，再通过孟德尔随机化等方法评估因果效应、eQTL和mQTL分析定量解释遗传影响、精细映射（fine-mapping）精准定位治病位点，最后，还可进行基因功能的实验验证并预测临床数据。以上本团队总结了当前中医复杂表型跨尺度关联研究，给出了中医表型组学的病-证研究思路[5]。但中医表型组学的数据分析和表型组学一样，需要基于足够可靠稳定的表型采集技术、足够规模的数据样本量、及够全面的多组学数据。当前，已有一批具有中医特色的队列研究正在如火如荼地开展[6]，中医表型组学研究方兴未艾，发展前景十分广阔。

3 总结与展望

中医表型组学作为一门新兴学科，融合了中医学、基因组学、计算机科学等多个学科特点，注重挖掘跨尺度表型的深层联系，进而推动中医理论现代化。本研究简述了中医表型组学的数据类型和数据特点，并简要说明了中医表型组学数据的处理方法和跨尺度研究分析方法。中医学通过大量的临床实践，将大量疾病表型执简驭繁地概括为若干具有特色的中医表型，而对于这些表型与现代生命科学的深层关联，则需要依靠表型组学的理念方法和大量的数据验证。中医学与表型组学是“道”与“术”的关系，二者在理念和技术上互为补充和借鉴。

就未来的研究方向而言，中医表型组学具有广阔空间：一是挖掘中医特色表型（例如舌象、脉象、证候等）的遗传结构特征与疾病机制；二是基于特殊中医表型特点，开发模式动物或技术平台；三是根据表型-药效关系，发现药物新靶点或设计新药；四是开发标准化中医表型测量方法及算法等；五是融合人工智能，实现疾病早期诊断或实现自我健康监测。

当前，中医表型组学尚处于学科早期建设阶段，仍面临许多挑战，例如表型的提取和数据规模就是当前亟待解决的两个问题。首先是表型量化的标准化，大量复杂的中医表型存在人工依赖性强、主观色彩浓的问题，如何将复杂、人工识读的证候等表型提炼归纳为客观的表型特征，当前还没有一个统一的解决方案；其次是数据规模，目前仍处于数据早期积累阶段，仍然缺乏中医表型的大样本队列研究，尤其是健康人群的队列研究。此外，数据规模、数据质量需要进一步提升，数据处理也仍需要形成规范。对于前者，将复杂的中医表型尽可能拆分成单一的、可重复、易采集的简单表型，可能是一种解决途径。例如，将阴虚证拆解为舌象、脉象、体温、基础代谢率、血尿便指标等多种客观表型的集合，再研究其遗传结构和机制关联。而对于后者，则依赖于多学科的交叉碰撞和更多研究者的积极探索，建立广泛的研究合作和数据共享，或形成开源的公共数据库。

中医学本身强调整体观念和辨证论治，关注复杂因素间的动态变化和互作关系，与精准医学有着天然的共同点。随着多组学的技术发展，人体各层次的物质水平变化得以量化，中医表型联动跨尺度、多维度的现代生物医学大数据成为可能。中医表型组学的发展，是融合了中医学理论与现代科技对人体复杂系统的一次重新审视，有望深入解读中医理论的科学内涵，并为精准医学的发展提供新的视角。

参 考 文 献

[1] 金力. 寻找健康“密钥”: 从基因组到表型组[J]. 科学, 2022, 74(1): 1-6, 69.

[2] DENNY J C, COLLINS F S. Precision medicine in 2030-seven ways to transform healthcare[J]. Cell, 2021(6): 1415-1419.

[3] 仝小林, 赵林华, 张博荀. 表型组学与中医药创新发展[J]. 科学, 2022, 74(1): 4, 22-25.

[4] 仝小林, 何莉莎, 赵林华. 助力精准医学: 表型组学与中医学的借鉴与融合[J]. 中国中西医结合杂志, 2018, 38(3): 368-370.

[5] 左玲, 陈建新, 王伟, 等. 中医信息学与表型组学: 症状的遗传突变与证候的生物学基础初探[J]. 北京中医药大学学报, 2022, 45(2): 140-147.

[6] 原淳淳, 王晶, 舒冰, 等. 中医表型组学的概念与相关研究体系的构建[J]. 中医杂志, 2022, 63(5): 407-411.

[7] 蔡轶珩, 胡绍斌, 关静, 等. 中医舌诊客观化技术发展分析及应用探讨[J]. 世界科学技术-中医药现代化, 2021, 23(7): 2447-2453.

[8] 韩鹏鹏, 王天芳, 廖结英, 等. 中医面部色泽望诊的客观量化研究进展[J]. 环球中医药, 2021, 14(4): 749-755.

[9] 夏雨墨, 高慧, 王庆盛, 等. 颜色空间在中医望诊客观化研究中的应用进展[J]. 中国中医药信息杂志, 2021, 28(4): 135-139.

[10] 胡洁娴, 刘旺华. 中医面部色诊现代研究进展[J]. 中医学报, 2020, 35(5): 1001-1005.

[11] 洪静, 陈聪, 许朝霞, 等. 中医声诊客观化研究进展[J]. 中华中医药杂志, 2019, 34(11): 5324-5326.

[12] ZHAO Z, ZUNG J L, HINZE A, et al. Mosquito brains encode unique features of human odour to drive host seeking. [J]. Nature, 2022, 605(7911): 706-712.

[13] 罗静静, 左晶晶, 季仲致, 等. 面向脉诊客观化的脉搏传感器研究综述[J]. 仪器仪表学报, 2021, 41(8): 1-14.

[14] 敖艺洲. 脉诊的客观化现代研究[J]. 实用中西医结合临床, 2021, 21(11): 158-159.

[15] JIN C, XIA C, ZHANG S, et al. A wearable combined wrist pulse measurement system using airbags for pressurization[J]. Sensors (basel), 2019, 19(2): 386.

[16] YAN J, CAI X, CHEN S, et al. Ensemble learning-based pulse signal recognition: classification model development study[J]. Jmir Med Inform, 2021, 9(10): e28039.

[17] 徐熊, 温川飙, 宋海贝. 浅谈中医脉诊信息化研究[J]. 成都中医药大学学报, 2020, 43(2): 55-59.

[18] 杜金行, 史载祥. 血瘀证中西医结合诊疗共识[J]. 中国中西医结合杂志, 2011, 31(6): 839-844.

[19] 孟昊, 尚菊菊, 胡元会, 等. 外周血不同大小微粒水平与冠心病及冠心病血瘀证患者的相关性[J]. 中医杂志, 2022, 63(7): 653-657.

[20] 许伟明, 胡镜清, 江丽杰. 当代病证结合研究思路和方法进展评析[J]. 世界科学技术-中医药现代化, 2016, 18(5): 769-775.

[21] 郑昊, 黄嘉华, 潘美均, 等. 近5年中医诊断量表文献分析[J]. 河南中医, 2020, 40(11): 1703-1708.

[22] 白明华, 李倩茹, 李竹青, 等. 中医体质量表的国内外研究概述[J]. 中华中医药杂志, 2021, 36(10): 5993-5996.

[23] 白明华, 王济, 郑燕飞, 等. 基于108015例样本数据的中国人群中医体质类型分布特征分析[J]. 北京中医药大学学报, 2020, 43(6): 498-507.

[24] 朱燕波, 王琦, 虞晓含, 等. 8 448例一般人群中医体质兼夹状况分析[J]. 中华中医药杂志, 2020, 35(9): 4337-4341.

[25] 王琦, 朱燕波. 中国一般人群中医体质流行病学调查—基于全国9省市21 948例流行病学调查数据[J]. 中华中医药杂志, 2009, 24(1): 7-12.

[26] 朱燕波, 王琦, 薛禾生, 等. 中医体质量表性能的初步评价[J]. 中国临床康复, 2006(3): 15-17.

[27] 朱燕波, 史会梅, 虞晓含. 不同条目版本的中医体质量表在健康人群中应用的性能比较[J]. 中国全科医学, 2019, 22(35): 4381-4387.

[28] 徐璡, 许朝霞, 王又闻, 等. 基于病证结合的中医证候规范化研究进展[J]. 中华中医药杂志, 2014, 29(1): 14-16.

[29] 谭婧泽, 彭倩倩, 李金喜, 等. 新疆维吾尔族牙齿形态特征及其与EDARV370A相关性的研究[J]. 中国科学: 生命科学, 2014, 44(1): 85-96.

[30] 张霄帅. 疾病遗传网络构建中两整体基因之间共关联作用的统计推断方法研究[D]. 山东大学, 2012.

[31] 李芳玉, 袁中尚, 张霄帅, 等. 基于偏最小二乘路径模型的整体基因与多数量性状交互作用的统计推断[J]. 山东大学学报(医学版), 2013, 51(11): 102-106.

[32] WU Y, ZHANG F, YANG K, et al. Symmap: an integrative database of traditional Chinese medicine enhanced by symptom mapping[J]. Nucleic Acids Research, 2019, 47(D1): 1110-1117.

[33] ZHANG Y, WANG N, DU X, et al. Sofda: an integrated web platform from syndrome ontology to network-based evaluation of disease-syndrome-formula associations for precision medicine[J]. Science Bulletin, 2022, 67(2095): 1097.

[34] QI Z, TU L, CHEN J, et al. The classification of tongue colors with standardized acquisition and icc profile correction in traditional Chinese medicine[J]. Biomed Res Int, 2016: 3510807.

[35] WANG Y, SHI X, LI L, et al. The impact of artificial intelligence on traditional Chinese medicine[J]. Am J Chin Med, 2021, 49(6): 1297-1314.

[36] 许家佗, 周昌乐, 方肇勤, 等. 舌像颜色特征的计算机分析与识别研究[J]. 上海中医药大学学报, 2004, 18(3): 43-47.

[37] 许家佗, 孙炀, 张志枫, 等. 基于差分统计方法的舌象纹理特征的分析与识别[J]. 上海中医药大学学报, 2003, 17(3): 55-58.

[38] WANG X, WANG X, LOU Y, et al. Constructing tongue coating recognition model using deep transfer learning to assist syndrome diagnosis and its potential in noninvasive ethnopharmacological evaluation[J]. J Ethnopharmacol, 2021: 114905.

[39] WANG X, LIU J, WU C, et al. Artificial intelligence in tongue diagnosis: using deep convolutional neural network for recognizing unhealthy tongue with tooth-mark [J]. Comput Struct Biotechnol J, 2020, 18: 973-980.