文章编号:1671-7554(2025)08-0001-16 DOI:10.6040/j.issn.1671-7554.0.2025.0568 ·专家述评·

AI 语言表征的多模态大数据队列设计理论方法体系

薛付忠[1,2,3]

(1.山东大学齐鲁医学院公共卫生学院医学数据学系, 山东 济南 250012;
2.国家健康医疗大数据研究院, 山东 济南 250003; 3.山东大学齐鲁医院, 山东 济南 250012)

摘要:本研究突破传统流行病学队列设计的理论框架,创立 AI 语言表征的多模态队列理论方法体系,形成 AI 语言建模的多模态队列新范式。该体系整合健康档案、电子病历、影像、基因等多源异构数据,借助 Transformer 等 AI 模型进行低维嵌入,统一量化为多模态嵌入向量。围绕"数字组学-数字生物标记-数字表型"三层架构,提出多模态融合、嵌入向量生成、因果推理等关键方法。创新性提出数字生物标记需满足 PICLS 准则:可预测性(predictable)、可解释性(interpretable)、可计算性(computable)、潜变量性(latent-variable)、稳定性(stable);数字表型在此基础上还应满足终点性(endpoints),即 PICLSE 准则,确保多模态队列的应用价值。技术方面,本文详述了嵌入生成、数据编码/解码、数据库构建及标记提取等流程。以猩红热主动监测为应用案例,展示多模态嵌入队列的实际应用效果。该体系为流行病学队列研究提供了新范式,对推动精准医疗与公共卫生智能化具有重要意义。

关键词:AI 语言表征;多模态队列;数字组学;数字生物标记;数字表型;PICLS/PICLSE 准则
中图分类号:R181.2+3 **文献标志码**:A

Theoretical and methodological framework for multimodal big data cohort design based on AI language representation

XUE Fuzhong[1,2,3]

(1. Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China; 2. National Institute of Health and Medical Big Data, Jinan 250003, Shandong, China; 3. Qilu Hospital of Shandong University, Jinan 250012, Shandong, China)

Abstract: This paper proposes a theoretical and methodological framework for multimodal cohort design based on artificial intelligence (AI) language representation, breaking through the conventional paradigm of traditional epidemiological cohort studies and establishing a novel model for language-based multimodal integration. The framework integrates heterogeneous medical data—such as health records, electronic medical records, medical imaging, and genomic information—into a unified low-dimensional embedding space using Transformer-based models. Centered on a three-layer architecture of "Digital Omics-Digital Biomarkers-Digital Phenotypes", it introduces key methods including embedding vector generation, causal inference, and multimodal data fusion. The study innovatively defines the PICLS criteria for digital biomarkers: predictability, interpretability, computability, latent-variable structure, and stability. On this basis, digital phenotypes are further required to meet the endpoints criterion, forming the PICLSE criteria to ensure their clinical utility in disease prediction and intervention. Technically, the paper details the entire process of embedding generation, data encoding/decoding, database construction, and biomarker extraction. A case study on scarlet fever surveillance demonstrates the practical application of the proposed multimodal embedded cohort in clinical screening and intelligent early warning. This framework offers a novel paradigm for epidemiological cohort research and provides methodological support for advancing precision medicine and smart public health.

Key words: AI language representation; Multimodal cohort; Digital omics; Digital biomarkers; Digital phenotypes; PICLS/PICLSE criteria

收稿日期:2025-05-16
基金项目:国家自然科学基金重点项目(82330108);国家自然科学基金面上项目(82173625);潍坊市中央财政支持公立医院改革与高质量发展示范项目(ZFCG-2024-0000505);河南省重大科技专项项目(241100310300)
通信作者:薛付忠。E-mail:xuefzh@ sdu.edu.cn

传统队列设计与分析方法在医学研究中曾取得辉煌成就，但在面对现代医学多源异构、动态复杂的大数据环境时，逐渐显现出多重局限性。首先，在信息维度方面存在显著限制。传统方法依赖结构化表格数据，难以有效整合和处理非结构化文本、图像数据及连续生命体征等高维复杂信息，导致大量有价值的信息无法被纳入分析框架。其次，传统统计方法面临“黑箱”困境。在生物医学系统中，机制复杂、变量众多、关系模糊，传统方法常基于线性假设或变量简化处理，难以揭示多层次变量之间的潜在因果路径，犹如在黑箱中操作，缺乏机制可解释性和预测稳定性。再次，因果推断能力薄弱是另一大瓶颈。由于真实世界中的因果链常伴随混杂因素和系统偏倚，传统统计方法即便采用前沿设计，仍难以完全消除这些干扰，造成推断结果的不确定性和外部效度不足。综上，传统队列方法已难以满足对高维、多模态、生物机制可解释性和真实因果推断的研究需求，迫切需要重构以 AI 语言表征驱动的新型队列设计与分析范式。因此，传统的流行病学队列研究在面对爆炸式增长的多源异构健康医疗大数据时，其既有的理论框架已难以适应。本文以前瞻性的科学视角，突破传统流行病学队列设计的理论桎梏，独辟蹊径地创立基于 AI 语言表征的多模态队列理论方法体系，从而开创基于 AI 语言建模的多模态队列研究全新范式。

这一创新性体系的核心在于其对海量健康医疗数据的整合与处理能力。它融合了包括健康档案、电子病历、医学影像以及生物传感器等多源异构数据。借助 Transformer 模型等先进的 AI 模型，将这些原本分散、复杂的数据巧妙地转化为统一的低维嵌入向量[1]，实现了不同模态数据在同一语义空间的高效对齐与深度集成。这种基于 AI 语言表征的嵌入技术[2-16]，是构建多模态队列的基石，它将医疗数据的复杂性转化为可计算、可分析的数字形式，为后续的智能应用奠定了坚实基础。

本文围绕“数字组学-数字生物标记-数字表型”这一独创性三层架构，系统阐述多模态融合、嵌入向量生成以及因果推理等一系列关键方法。尤为重要的是，本研究提出数字生物标记必须满足的 PICLS 严苛验证准则，即具备可预测性（predictable）、可解释性（interpretable）、可计算性（computable）、潜变量性（latent-variable）和稳定性（stable）。在此基础上，为数字表型进一步增加了终点性（endpoints）要求，形成更为完善的 PICLSE 准则，从而有力保障多模态队列在疾病预测和干预中的实际应用价值和临床意义。为了确保这些数字生物标记和数字表型的科学性和可靠性，创新性地引入了严谨的因果推理框架，通过相关性分析、干预效应评估、反事实推理[17]等多层次、多维度的分析，确保所提取的数字特征不仅与健康结局存在统计关联，更具备确切的因果解释力[17]。这标志着从传统的统计学关联分析迈向了更深层次的因果关系挖掘，极大提升了研究结论的医学可信度和指导意义。

在技术方法实现层面，本文详细剖析了 AI 语言表征的多模态队列的设计方法，包括多模态数据输入、数据编码、嵌入向量生成、数据解码与复原的全流程，并构建了支持这些操作的数字组学数据库[18-20]。数字组学的概念在本体系中被赋予了全新的内涵，它不再仅仅是数据的简单汇集，更是通过 AI/ML 技术的赋能，实现个体化诊疗的新兴范式。数字生物标记[1]和数字表型[21]的提取与验证方法在本体系中得到了系统性的构建，特别结合因子分析、结构方程模型[22]以及多环境因果推理[17]等方法，确保了这些数字标记的潜变量性和稳定性。同时，通过 Prentice 准则和因果中介分析[23-24]等手段对数字表型的终点性进行严格验证，使其能够作为可靠的临床替代终点，加速临床研究进程。

本文以猩红热主动监测为例，展示 AI 语言表征的多模态队列在实际应用中的巨大潜力。针对传统传染病监测系统中早期漏诊导致疫情扩散的难题，研究团队构建了一个基于巢式病例对照设计的多模态数据库，并在此基础上，巧妙地利用 LoRA 微调开源大模型[8-9]，成功开发了猩红热主动监测 AI 大模型。该模型在多中心、多模态数据上的交叉验证表现出了令人瞩目的性能，尤其是在仅使用门诊病历的情况下也能获得良好的监测效果，极大地提高了猩红热早期预警和防控能力，为遏制疫情蔓延提供强大的技术支撑。

本研究工作的意义不仅在于为流行病学队列研究提供了崭新的范式，更为推动精准医疗和公共卫生智能化提供了强有力的理论与方法支持。随着技术的不断演进，AI 语言表征的多模态队列有望在疾病的精准预测、个性化诊疗以及智能健康管理等领域带来变革性的突破，实现从海量多模态数据中挖掘深层价值，驱动未来医学走向智能化、精准化。

1 基本概念

Transformer 通过自注意力机制高效建模长文本语义，推动了大规模预训练语言模型的发展。近年来，医学领域涌现出一系列特定模型，如生物医学语言预训练模型（bidirectional encoder representa-

tions from Transformers for biomedical text mining, BioBERT)[2]在通用基于 Transformer 的双向编码器(bidirectional encoder representations from Transformers, BERT)基础上引入生物医学文献语料进行预训练,显著提升命名实体识别等任务表现;临床语言预训练模型(clinical notes using bidirectional encoder representations from Transformers, ClinicalBERT)[3]使用临床笔记语料微调,使模型更适应临床场景;生物医学语言理解与推理模型(the biomedical language understanding evaluation bidirectional encoder representations from Transformers, BlueBERT)[4]、基于 PubMed 文本的预训练语言模型(a BERT model pre-trained on PubMed text, PubMedBERT)[5]和大规模生物医学领域语言模型(larger biomedical domain language model, BioMegatron)[6]等进一步扩大数据和参数规模,增强了医学语言理解能力。这些模型为 Transformer 构架下的 AI 语言表征的多模态队列创建提供了理论方法基础。随着算力提升,医学专用语言模型不断扩展。Yang 等[7]构建的 GatorTron 基于 900 亿词大语料训练三种规模模型,在多个临床任务中性能领先,并进一步发展出生成模型 GatorTronGPT[8],Google 推出的 Med-PaLM[9]、Med-PaLM 2[9]等系列模型通过指令微调强化医学问答能力。更重要的是,多模态学习将图像、音视频等与文本嵌入统一语义空间,拓展医学语言表征能力。OpenAI 提出的对比语言-图像预训练(contrastive language-image pre-training, CLIP)采用对比学习联合训练图文模型,在通用任务中表现卓越[10];受其启发,MedCLIP[12]利用未配对医学图文数据结合对比策略,在医学检索等任务中超越传统方法;MoveNet[12]、VGGish[13]与 VGGFace[14]分别从动作、语音和面部表情提取高层语义特征,形成低维但表达力强的嵌入向量。微软的 BioViL[15]融合图像与影像报告实现细粒度对齐,在术语归一等任务中取得良好表现[16]。DeepMind 的 Flamingo[25]通过少样本跨模态注意力机制,在多模态问答等 16 项任务上表现领先。上述研究展现了 Transformer 编码在 AI 语言表征的多模态队列创建方面的巨大潜力。

基于以上研究进展,本文提出 AI 语言表征的多模态队列的基本概念:以 Transformer 为基础,融合健康档案、电子病历、影像、基因、传感数据等多源异构健康医疗数据,通过深度学习技术统一转化为低维嵌入向量,实现不同模态数据在同一语义空间的对齐与集成,以支持数据标准化、嵌入生成、语义增强与向量索引,用于疾病预测、临床研究、临床决策、药物研发等领域,推动个性化、智能化数字医学发展,构建医学人工智能研究的队列数据库与分析框架。见图 1。

图 1 AI 语言表征的多模态队列设计概念框架

Figure 1 Conceptual framework for the multimodal queue design of AI language representation

如图 1 所示,AI 语言表征是将复杂异构的多模态数据转化为低维度向量,以便于能够在同一空间内表征和处理不同类型的数据。通过上述 Transformer 架构与多模态学习[2-16],将不同类型的数据

映射到同一低维空间内，从而使得相关数据在嵌入空间中更加接近，有助于进一步的分析和预测。在其设计框架中，收集并标准化处理来自不同源的异构数据，形成多模态数据流。这些数据通过深度嵌入模型转换为嵌入向量，模型提取数据的关键信息并进行低维度表示，保留数据的主要特征，形成支持增强检索的索引。生成的嵌入向量可用于后续的分析任务，如疾病预测和临床决策支持等。通过这些嵌入向量，AI模型能够实现病因精准推断、健康风险评估、疾病进展监测和精准诊疗。嵌入向量不仅是简化的数据表示，还包含了数据中潜在的规律，支持跨模态的医疗决策。这一过程结合了深度学习的优势，能够处理来自不同来源的异构数据，提供个性化的医疗服务。最终，AI语言表征的多模态队列分析为临床决策、临床试验、药物研发等应用提供了强大的多模态向量化数据支持。

2 基本原理

多模态嵌入理论方法是构建AI语言表征的多模态队列的核心。例如，上述Transformer多模态学习[2-16]架构下，如何实现高效的向量嵌入，就是构建多模态队列的关键。近年来，将文本、影像、基因等多模态数据映射至统一空间的理论方法有了长足的发展。例如，CLIP等[10,26]模型采用对比学习对齐图文表示，在医学中也用于病理图像与分子特征的关联推断。在融合架构方面，MADDi框架[27]引入跨模态注意力机制整合磁共振成像（magnetic resonance imaging，MRI）、基因和临床数据，实现96.9%诊断准确率，展现强大协同效应；Flamingo等模型[28]启发医学领域利用中间融合策略，实现深度跨模态语义建模。又如，在基因组数据嵌入中，将脱氧核糖核酸（deoxyribonucleic acid，DNA）视作“语言”的多基因编码器，实现了基因信息与语言模型融合，提升疾病预测能力[29]；此外，多模态嵌入对齐图文、基因等特征，构建统一语义空间，使模型可跨模态推理并生成诊断报告[16]。

基于上述诸多进展，本文提出AI语言表征的多模态队列的基本原理：以基于Transformer深度学习模型的多模态数据AI语言表征多模态队列为例，强调了数字生物标记、数字表型和数字组学等关键理论方法。见图2。

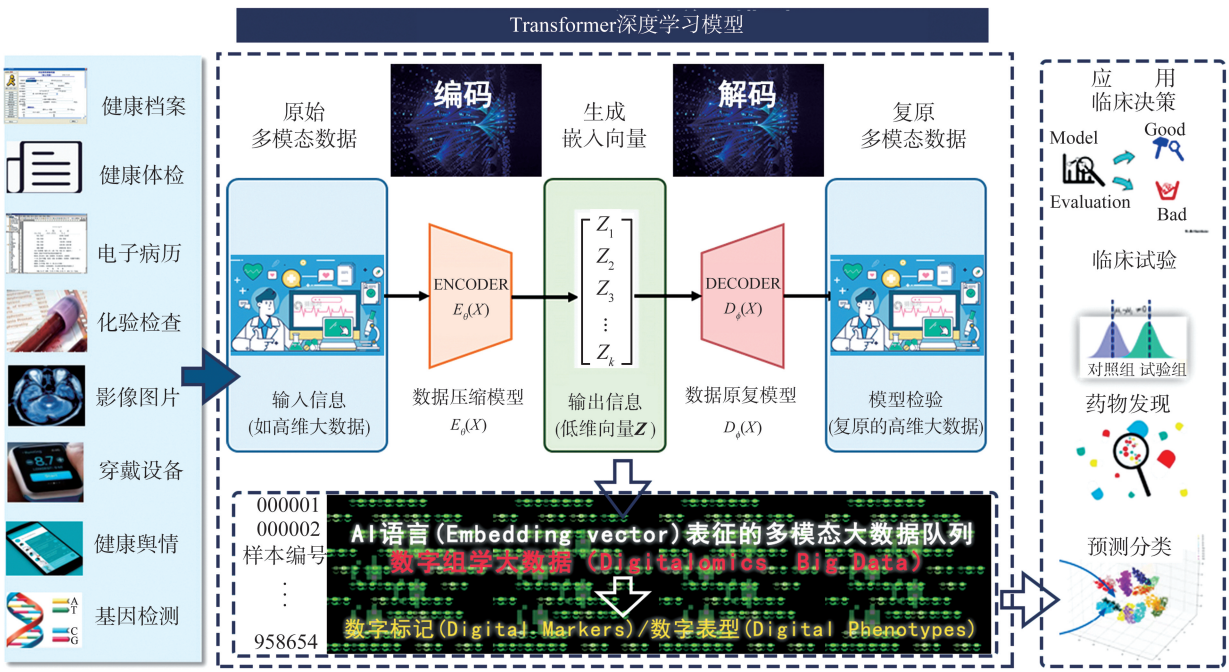


图2 AI语言表征的多模态队列设计基本原理

Figure 2 Basic principles of multimodal queue design for AI language representation

首先，将多源多模态数据（如健康档案、电子病历、体检数据、影像数据等）进行收集、链接、映射和融合。这些数据通过编码器转化为低维度的嵌入向量，即将原始的高维数据表示转化为统一的低维数字化向量。这些嵌入向量是对多模态数据的数字化表达，它们包含了数据的关键信息，便于深度学习模型进一步处理。

随后，解码器将低维嵌入向量解码，重构多模态数据，或生成预测数据，例如预测疾病风险或临床诊断。这样，经过处理的低维嵌入向量会被解码器转

化为可解释的输出。例如,用于识别数字表型、预测疾病风险、构建临床决策模型等。这一过程中的数字表型表现为具体的、通过数据转化得到的健康或疾病特征,能够为个性化医疗和疾病预测提供有价值的信息。

多模态数据以 AI 语言表征的数字组学呈现。从数字组学中,提取数字标记,量化人体健康的各种特征。进一步提炼为表征具体的疾病终点结局的数字表型。通过 AI 语言的数字组学表征,将数字生物标记和数字表型高效提取并转化为具有流行病学意义和临床意义的特征标记。在 AI 语言表征的多模态队列中,设计如下几方面的核心原理。

2.1 数字组学

数字组学旨在将生物组学、影像、生理信号、可穿戴设备和电子病历等多模态数据融合,通过 AI/ML 实现个体化诊疗的新兴范式[18],其核心在于多模态信息整合和数据驱动预测,广泛应用于心血管疾病风险评估、数字听诊(phonocardiogram, PCG)和大型语言模型辅助诊断等场景。数字转型技术(如高性能计算与张量建模)加速从数字组学中发现数字生物标志物,有效支持精准治疗策略,尤其在癌症和心血管等领域展现应用前景[19]。此外,数字孪生新概念通过建模个体特征助力靶向干预,推动数字化干预与远程医疗深度融合,预示未来心血管医学的精准与个体化转型[20]。

基于上述背景,本文提出数字组学的基本概念和原理:数字组学是一种跨学科概念,融合数据科学、计算机科学、工程学、统计学、物联网和人工智能等技术,以标准化方式集成和管理多模态数据。其核心目标是将健康档案、体检数据、生物传感器数据、基因检测、影像数据等多源异构数据进行数字化、无损压缩和整合,以创建高效、可分析的 AI 语言表征向量化数据体系。其基本原理:通过物联网传感器、生物组学、个人健康记录(personal health record, PHR)和电子健康档案(electronic health record, EHR)将多模态数据链接、映射和融合,并通过 AI 模型进行处理。AI 模型在这一过程中负责数据预处理、特征提取和深度嵌入,以便后续将高维数据转化为低维的数字表型和数字生物标记,形成标准化的数字组学队列。见图 3。

图 3 AI 语言表征的多模态数字组学

Figure 3 Multimodal digitalomics for AI language representation

AI 语言表征是数字组学的核心技术,它通过深度学习方法将多模态数据转化为统一的低维度向量表示,使得不同类型的数据能够在同一计算框架下进行分析和推理[2-16,25-29]。这一过程中,数字组学通过数据融合、压缩和编码-解码过程,确保数据的完整性和有效性,并提升 AI 模型在健康预测、个性化医疗、精准预防等领域的应用价值。最终,数字组学支持在线实时精准预防、智能健康管理和疾病预测。通过 AI 模型的赋能,它能够挖掘健康数据中的深层信息,优化临床决策。

2.2 数字生物标记和数字表型

目前,数字生物标记与数字表型的概念仍不完

善。Sameh 等[21]认为数字生物标记是通过数字设备客观获取的、可量化的生理或行为数据，用以指示生物学过程、疾病状态或对干预的响应；而数字表型指从个人数字设备收集的数据中提取的个体表现特征（可观察的身体或行为特征）。二者之间的关系：数字生物标记是数字表型中具有生物学指示意义的特征指标。Oudin 等[30]认为数字表型是通过个人数字设备对个体行为和生理状态进行持续、细粒度量化，是可以量化且与健康结果相关的具体指标。本文将数字生物标记和数字表型的定义及二者之间的关系阐述如下。

2.2.1　数字生物标记

2.2.1.1　数字生物标记的定义

数字生物标记是通过多模态数据采集、链接、映射、融合的方式，利用大数据赋能 AI 模型驱动的数字组学技术，从个体（群体）的多源、多维、多模态生物数据中提取出的客观、可量化的指标。这些指标反映了生理状态、病理进展或对治疗干预的反应，为精准医学、精准公共卫生学和个性化健康管理提供新型生物标记。

2.2.1.2　数字生物标记的验证

与传统生物标记类似，数字生物标记的应用需要严格的验证。目前，仅简单地将数字生物标记的验证分为技术验证（验证提取的方法可靠有效）和临床验证（验证该标记与临床结局的相关性和预测能力），尚缺乏严格的验证准则[20-21]。为此，如图 4 所示，本研究提出，数字生物标记必须满足如下 PICLS 准则：

（1）可预测性。数字生物标记需具备高预测能力，用于疾病筛查、风险评估和个性化治疗，这是对数字生物标记的最基本要求。即 AI 模型结合多模态数据（生理信号、基因组、影像）所提取的数字生物标记最起码具备预测准确性，如心率变异性可预测心脏病风险。

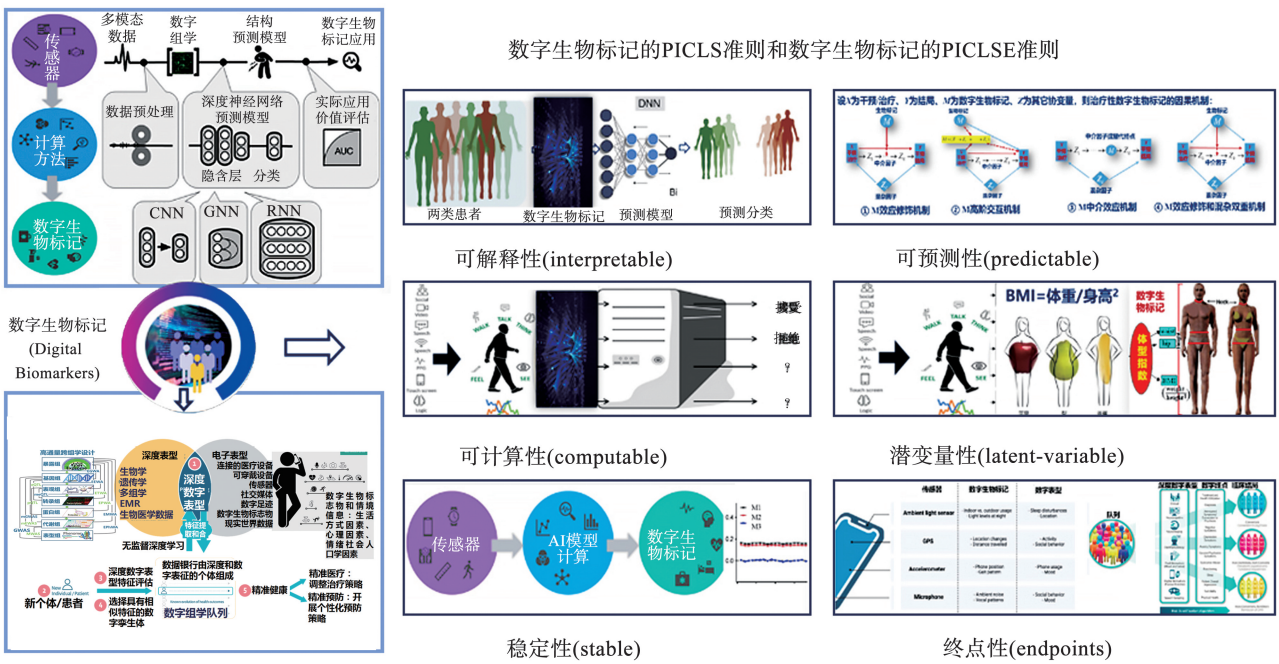


图 4　数字生物标记与数字模型及其准则

Figure 4　Digital biomarkers, digital models, and criteria

（2）可解释性。数字生物标记应具有因果解释能力，提高医学可信度。例如，通过因果推断分析 BMI 如何影响代谢疾病。采用 SHAP、LIME 等方法提高 AI 决策透明度，助力精准医学。

（3）可计算性。数字生物标记应能通过计算机算法高效处理，如智能感知系统采集数据（语音、步态、心率），经 AI 分析生成健康特征。数字孪生技术结合可计算生物标记，优化疾病管理。

（4）潜变量性。数字生物标记无法直接测量，只能由可测变量整合为潜变量，例如，由可测指标体质量和身高计算 BMI，就是最简单的例子。更复杂的情形，通过因子分析、潜变量建模揭示疾病隐含特征，如认知障碍的多维表型分析。

（5）稳定性。数字生物标记应在不同环境、设备和个体间保持一致性，如远程监测中智能手表和医疗设备的血压数据需一致。采用标准化信号处理和纵向研究验证数据可靠性。

2.2.2　数字表型

2.2.2.1　数字表型的定义

数字表型是指利用数字/AI 技术，将个体或群

体多源多模态数据转化为数字组学并提取出的客观、可量化数字生物标记;进而,对应疾病组学循证知识图谱[31],所定义的具有生物学意义和临床价值的数字终点。这些数字终点反映特定健康状态或疾病进展(如帕金森病运动症状、心力衰竭早期迹象等),提供对生理状态、病理进展及其对干预措施反应的深度洞察。

2.2.2.2 数字表型的验证

由于对数字表型的定义[32]缺乏严格的因果推理理论,目前尚缺乏数字表型验证的准则,为此,如图 4 所示,本文提出数字表型必须满足 PICLSE 准则:即除了满足上述数字生物标记的可预测性、可解释性、可计算性、潜变量性和稳定性以外,还必须满足终点性。即,数字表型的最终目标是表达临床终点,如预测某种治疗能否改善生存率。数字表型在临床研究中可作为替代终点,缩短新药研发等时间,提高效率。

3 技术方法

3.1 AI 语言表征的多模态队列设计方法

针对如图 2 所示的 AI 语言表征的多模态队列设计基本原理,基于 Transformer 深度学习模型[33],通过向量嵌入[2-16,25-29],进行多模态数据 AI 语言表征的技术方法及技术流程如下。

3.1.1 多模态数据输入

设多源多模态数据 X 由 N 个数据模态组成:$X=\{X_1,X_2,\cdots,X_N\}$,其中,每个 X_i 代表不同来源的数据,如:健康档案 X_H、EHR X_E、体检数据 X_T、影像数据 X_I、基因检测 X_G 等。每个数据模态 X_i 由 d_i 维特征组成:

$$X_i \in \mathbf{R}^{n_i \times d_i}, \quad i=1,2,\cdots,N, \tag{1}$$

其中,n_i 是样本数,d_i 是模态 X_i 的特征维度。

为了将不同数据模态映射到统一的特征空间,需进行数据链接、映射与融合:

$$\boldsymbol{X}'=F(X)\in\mathbf{R}^{n\times d}, \tag{2}$$

其中,$F(X)$ 是多模态特征融合函数(如拼接、注意力机制等),$\boldsymbol{X}'$是融合后的特征矩阵,d 是最终对齐的特征维度。

3.1.2 数据编码

在 Transformer 编码器中,数据 $\boldsymbol{X}'$通过线性变换和多头注意力机制被转换为嵌入表示:

$$\boldsymbol{Z}=E_\theta(\boldsymbol{X}'), \tag{3}$$

其中,E_θ 是 Transformer 编码器。$\boldsymbol{Z}$ 是嵌入向量,$\boldsymbol{Z}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}$,其中 $\boldsymbol{Q}=\boldsymbol{X}'\ \boldsymbol{W}_Q$ 是查询矩阵,$\boldsymbol{K}=\boldsymbol{X}'\boldsymbol{W}_K$ 为键矩阵,$\boldsymbol{V}=\boldsymbol{X}'\boldsymbol{W}_V$ 为值矩阵,d_k 为 Transformer 计算的维度。最终得到低维嵌入向量:$\boldsymbol{Z}=\{z_1,z_2,\cdots,z_K\}$,$\boldsymbol{Z}\in\mathbf{R}^{K\times d}$,其中,$K$ 是 Transformer 生成的序列长度,d 是嵌入空间的维度。

3.1.3 生成嵌入向量

编码器输出的嵌入向量 $\boldsymbol{Z}$ 代表了多模态数据的低维特征表示:

$$\boldsymbol{Z}=E_\theta(\boldsymbol{X}'), \tag{4}$$

其中,嵌入向量可用于疾病预测 $P(y\mid\boldsymbol{Z})=\mathrm{Softmax}(W\boldsymbol{Z}+b)$,其中,$y$ 是预测的疾病类别,W 和 b 是全连接层参数;个性化干预方案相似性匹配 $\mathrm{Sim}(z_i,z_j)=\left(\frac{z_i z_j}{(\|z_i\|\ \|z_j\|)}\right)$,即计算不同患者之间的相似度。

3.1.4 数据解码

解码器用于将低维嵌入向量 $\boldsymbol{Z}$ 依概率恢复为原始的多模态数据,或者生成预测数据:

$$\hat{X}=D_\varphi(\boldsymbol{Z}), \tag{5}$$

其中,D_φ 是 Transformer 解码器,$\hat{X}$ 是重构(估计)后的数据。解码过程通常使用自回归方法 $\hat{x}_t=f(\hat{x}_{t-1},z_t)$,其中 $\mathrm{Fusion}(\cdot)$ 是解码器的非线性变换(如 LSTM、Transformer 解码器)。最终,解码器生成的输出重构原始数据 $\hat{X}\approx\boldsymbol{X}'$,或生成预测数据(如未来健康状态预测等)。

3.1.5 多模态数据复原

解码后的数据 $\hat{X}$ 经过映射,重新转换回多模态数据空间:

$$\hat{X}_i=G(\hat{X}),\ \hat{X}_i\in\mathbf{R}^{n_i\times d_i}, \tag{6}$$

其中,$G(\hat{X})$ 是映射函数,用于还原各个模态的数据。这一步确保 AI 语言表征生成的结果可以应用于健康监测、疾病筛查、个性化医疗等领域。

3.1.6 AI 语言表征的转化应用

转换后的多模态数据 $\hat{X}$ 可用多种场景,例如:

(1) 临床健康监测

$$R=W_R\hat{X}+b_R, \tag{7}$$

其中,R 代表健康风险评分。

(2) 个性化诊疗推荐

$$T=\mathrm{argmax}(W_T R+b_T), \tag{8}$$

其中,T 代表最佳治疗方案(如手术、靶向治疗)。

(3) 疾病预测

$$P(D\mid\hat{X})=\mathrm{Softmax}(W_D\hat{X}+b_D), \tag{9}$$

其中,$P(D)$是疾病分类的概率分布。

3.2 数字组学数据库构建方法

构建数字组学数据库[18-20]，需从多模态数据采集、数据标准化、数据融合、向量化存储、嵌入式表征等核心步骤进行建模。

3.2.1 多模态数据采集、链接与映射

数字组学数据库整合多个数据源，包括传感器数据 S、基因组数据 G、影像数据 I、EHR E、PHR P 等。进而，将不同来源的数据映射到统一的数据集：

$$D=\bigcup_{i=1}^{N}\{S_i,G_i,I_i,E_i,P_i\}, \tag{10}$$

其中，D 代表数字组学数据库，N 代表数据/个体数目。对于每种数据类型，定义数据标准化映射 $f:X\to\tilde{X}$。

3.2.2 多模态数据融合

多模态数据需要通过特定的数据对齐和融合策略进行整合。例如，采用联合表示学习[34]进行数据融合：

$$Z=\text{Fusion}(S,G,I,E,P), \tag{11}$$

其中，Fusion（·）代表数据融合函数。常见方法包括：数据拼接（concatenation）$Z=[S,G,I,E,P]$；注意力加权（attention-weighted fusion）$Z=\sum_{i=1}^{M}\alpha_i X_i$，$\sum_{i=1}^{M}\alpha_i=1$，其中，$\alpha_i$ 为注意力权重，M 为数据模态数。

3.2.3 数据向量化存储

为了高效存储，需对融合后的数据 Z 进行向量化编码：

$$\boldsymbol{V}=\varphi(\boldsymbol{W}Z+b), \tag{12}$$

其中，$\boldsymbol{W}$ 是转换矩阵，b 偏置项，$\varphi(\cdot)$ 是非线性激活函数（如 ReLU 或 Sigmoid）。这些向量 $\boldsymbol{V}$ 组成了数据库的核心存储格式：

$$D=\{\boldsymbol{V}_1,\boldsymbol{V}_2,\cdots,\boldsymbol{V}_N\}, \tag{13}$$

其中，D 代表数字组学数据库，$\boldsymbol{V}_i$ 是第 i 个个体的特征向量。

3.2.4 AI 语言表征的嵌入

为了优化数据库，以便于查询或分析，需采用深度嵌入方法[2-9, 10-29]，将原始数据映射到低维空间：$E=\text{Embedding}(\boldsymbol{V})$。例如，使用 Transformer 进行表征 $E=\text{Transformer}(\boldsymbol{V})$，其计算方法为：

$$h_i=\text{SelfAttention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i)=h_i=\text{softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_i, \tag{14}$$

其中，$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 分别是查询、键、值矩阵，d_k 是向量维度。

3.2.5 AI 驱动的数据库查询与分析

基于存储的数字组学数据，可以进行疾病预测、个性化医疗分析、健康风险评估等查询和应用。例如，疾病预测 $Y=f(\text{DM},\text{DP})+\grave{o}$，其中：$f(\cdot)$ 是 AI 预测模型（如 LSTM、Transformer 或 GNN），$\grave{o}$ 是噪声项；预测模型收敛阈值由损失函数（如交叉熵损失）L 确定，计算公式：

$$L=-\sum_{i=1}^{N}y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)。 \tag{15}$$

3.3 数字生物标记及数字表型提取与验证方法

上述数字组学数据库 D 是一个多模态融合、向量化存储、AI 语言表征的数据系统，从该数据库中提取数字生物标记并验证其 PICLS 准则；进而，提取数字表型及验证 PICLSE 准则的技术方法如下。

3.3.1 数字生物标记提取

从融合后的数字组学数据库 D 提取出的数字生物标记可表示为：

$$\text{DM}=\{\text{dm}_1,\text{dm}_2,\cdots,\text{dm}_k\},\ \text{dm}_i\in\mathbf{R}, \tag{16}$$

其中，DM 是数字生物标记集合，用于描述个体的生理或病理特征，k 是数字生物标记的数量，每个 dm_i 代表一个经过优化选择的生物特征，如心率变异性、血糖水平、基因表达特征等。提取数字生物标记的方法步骤为：

（1）低维特征提取。首先，通过深度学习方法从 Z 提取特征 F：

$$F=g_\varphi(Z), \tag{17}$$

其中，g_φ 是一个特征提取网络，如自动编码器（autoencoder，AE）等，$F\in\mathbf{R}^{N\times k}$ 是从融合数据库中提取出的数字生物标记。

（2）数字生物标记筛选。为了确保提取出的数字生物标记具有统计显著性和生物学意义，需采用主成分分析（principal component analysis，PCA）或特征选择算法（如 LASSO）：

$$\boldsymbol{W}^*=\arg\min_{\boldsymbol{W}}\|FW-Y\|^2+\lambda\|\boldsymbol{W}\|_1, \tag{18}$$

其中，$\boldsymbol{W}^*$ 是优化选择的特征权重矩阵，F 是目标变量（如疾病状态），λ 是正则化参数，用于控制特征选择的稀疏性。最终，选取权重 $\boldsymbol{W}^*$ 最高的 k 个特征作为数字生物标记 DM。

3.3.2 数字表型提取

数字表型是通过数字生物标记进一步计算得出的具有临床终点意义的高级生物特征 $\text{DP}=h(\text{DM})=\{\text{dp}_1,\text{dp}_2,\cdots,\text{dp}_m\}$，$\text{dp}_j\in\mathbf{R}$，其中 DP 是数字表型集合，$m$ 是数字表型的数量，每个 dp_i 代表一个健康结局状态的量化指标（如代谢综合征指数、帕金森病运动症状评分等）。提取数字表型的技术方法

如下：

（1）潜变量建模。数字表型无法直接测量，需要通过因子分析（factor analysis，FA）等潜变量方法进行建模 $X=\boldsymbol{L}F+\dot{o}$，其中，$X$ 是上述提取到的数字生物标记 DM，$\boldsymbol{L}$ 是因子载荷矩阵，F 是潜变量（即数字表型 DP，$\dot{o}$ 是噪声项）。

（2）数字表型优化。通过最大似然估计（maximum likelihood estimation，MLE）$L^{*}=\text{argmax}_{L}P(X|\boldsymbol{L})$ 优化后，最终得到潜变量数字表型 DP。

3.3.3 数字生物标记与数字表型的验证

如图 4 所示，数字生物标记必须满足 PICLS 准则：可预测性、可解释性、可计算性、潜变量性和稳定性；而数字表型除了满足 PICLS 准则以外，还必须满足终点性，即必须满足 PICLSE 准则。

为了确保数字生物标记和数字表型的可靠性，需要采用因果推理进行验证。利用因果推理分析数字生物标记 DM 与数字表型 DP 之间的因果关系，主要涉及三个层次[17]：① $P(Y|X)$ 相关性分析，基于观察数据计算概率关系；② $P(Y|\text{do}(X))$ 干预效应，评估 X 对 Y 的真实影响；③ 反事实推理，构建不同情境下的因果效应估计。其中，因果推理的核心目标是估计 $P(\text{DP}|\text{do}(\text{DM}))$，即，当主动干预数字生物标记 DM 时，干预对数字表型 DP 的影响。基于因果推理验证 PICLSE 准则的方法如下：

3.3.3.1 可预测性：因果预测 *vs.* 统计预测

传统预测模型仅使用相关性进行表型（疾病）预测 $\hat{Y}=f(\text{DM})+\dot{o}$；然而，因果推理要求计算干预后的因果效应 $P(Y|\text{do}(\text{DM}))$。具体方法可使用因果图（causal graph，G）识别混杂因子[17]：

$$P(Y|\text{do}(\text{DM}))=\sum_{C}P(Y|\text{DM},C)P(C), \tag{19}$$

其中，C 是混杂变量（如年龄、生活习惯）。

此外，也可采用工具变量法（instrumental variable，IV）进行因果推断 $Y=\alpha\text{DM}+\beta Z+\dot{o}$，其中，$Z$ 是工具变量，确保因果预测因子效应的估计。

3.3.3.2 可解释性：因果推理验证因果关系

因果解释的核心包括因果效应计算、因果路径分析、因果效应修饰分析。分述如下：

（1）因果效应计算。使用结构因果模型（structural causal model，SCM）[17] 计算因果效应 $\text{DP}=g(\text{DM},U)$，$U\sim P(U)$，其中 U 是不可观测的潜在混杂变量。$g(\cdot)$ 是因果作用函数。例如，基于因果推断 do 算子的因果效应计算方法为

$$E[\text{DP}|\text{do}(\text{DM})]=\int g(\text{DM},U)P(U)\text{d}U。 \tag{20}$$

（2）因果路径分析。数字生物标记 DM 可能通过中介变量 M 影响数字表型 DP：

$$P(\text{DP}|\text{do}(\text{DM}))=\sum_{M}P(\text{DP}|M)P(M|\text{do}(\text{DM}))。 \tag{21}$$

进而，通过 SHAP 计算因果贡献：

$$\varphi_i=\sum_{S\subseteq N\setminus\{i\}}\frac{|S|!\ (|N|-|S|-1)!}{|N|!}[f(S\cup\{i\})-f(S)], \tag{22}$$

计算每个数字生物标记 DM 在表型终点（疾病）中的因果贡献。其中，φ_i 代表生物标记 DM 对表型（疾病）Y 的因果贡献。这样，通过 do 算子可以计算生物标记的真实因果影响，而不是相关性影响。具体方法是通过分离直接效应与间接效应，识别中介因子。当数字生物标记 DM 通过中介变量 M 影响数字表型 DP 时，对上述总因果效应进行分解的方法如下。

总因果效应（total causal effect，TCE）：

$$\text{TCE}=P(\text{DP}|\text{do}(\text{DM}=1))-P(\text{DP}|\text{do}(\text{DM}=0))。 \tag{23}$$

自然直接效应（natural direct effect，NDE）：

$$\text{NDE}=P(\text{DP}|\text{do}(\text{DM}=1),M=M_0)-P(\text{DP}|\text{do}(\text{DM}=0),M=M_0), \tag{24}$$

其中，M_0 代表在基线水平下的中介变量状态。直接效应衡量 DM 直接影响 DP 的部分，不依赖中介变量 M。

自然间接效应（natural indirect effect，NIE）：

$$\text{NIE}=P(\text{DP}|\text{do}(\text{DM}=0),M=M_1)-P(\text{DP}|\text{do}(\text{DM}=0),M=M_0), \tag{25}$$

其中，M_1 代表在 DM = 1 时的中介变量状态。间接效应衡量 DM 通过 M 影响 DP 的部分。

这样，总效应可拆解为：TCE = NDE + NIE。

（3）因果效应修饰分析。在不同亚组人群或环境变量 Z 下，因果效应可能发生修饰，此时需要评估其依赖关系。引入修饰变量 Z 进行因果效应分析：

$$\text{DP}=\beta_0+\beta_1\text{DM}+\beta_2 Z+\beta_3(\text{DM}\times Z)+\dot{o}, \tag{26}$$

其中，Z 是可能修饰 DM→DP 因果关系的因子（如性别、年龄、基因型）；β_3 表示交互项的影响。如果 $\beta_3\neq 0$，则 Z 修饰了 DM→DP 的因果效应；如果 $\beta_3>0$，表示 Z 增强了 DM→DP 因果效应；如果 $\beta_3<0$ 表示 Z 减弱了 DM→DP 因果效应。

3.3.3.3 可计算性：因果估计的计算复杂度

在数字生物标记和数字表型的提取和评估过程中，通常需要进行大规模数据分析、特征选择、因果

推理和机器学习等海量计算。计算复杂度 $O(n\log n)$ 来源如下：

（1）低维特征提取。例如，采用 PCA 或特征选择（LASSO 回归）提取关键数字生物标记 $Z=WX$ 中，PCA 的计算复杂度（使用 SVD 分解）为 $O(nd^2)$，其中 d 是特征维度；由于实际数据通常高维，可以采用近似计算方法（如随机 SVD），降维可优化为 $O(n\log n)$。

（2）相关性分析。计算生物标记与表型的相关性。计算皮尔逊相关系数/互信息是的计算复杂度 $O(n)$；在高维数据上需排序操作，复杂度可达 $O(n\log n)$。

（3）因果推理。计算 $P(\mathrm{DP}|\mathrm{do}(\mathrm{DM}))$ 通常采用有向无环图（directed acyclic graph，DAG）建模。此时，需要对所有可能的因果关系进行搜索，采用贪心搜索或贝叶斯优化，时间复杂度通常为：$O(n\log n)$；在进行因果 SHAP 计算时，需要计算所有变量子集的贡献，最优情况下可用蒙特卡洛近似，时间复杂度 $O(n\log n)$。

（4）机器学习预测。生数字生物记对数字表型的预测。例如，梯度提升树 XGBoost 计算复杂度 $O(n\log n)$。

因此，计算复杂度最优情况下约为 $O(f(\mathrm{DM},\mathrm{DP}))=4O(n\log n)$。

3.3.3.4 潜变量性：因果推理识别不可观测变量

潜变量性是数字生物标记和数字表型需要具备的一项重要准则。它指的是数字生物标记无法直接测量，只能通过多个可测变量的组合推断出来。数字表型往往由多个潜在特征决定，需要通过统计或机器学习方法估计。例如，BMI 是由身高和体质量计算得出的，而代谢健康指数可能由多种血液生物标志物共同决定。

理论上，潜变量 L 和可观测数据 X 之间的关系可以表示为：$X=LF+\dot{o}$，其中，X 是观测变量（如数字组学向量的元素等），L 是潜变量（如代谢健康指数，或认知能力），是因子载荷矩阵，$\dot{o}$ 是误差项。验证的目的是确定否存在潜在结构变量 L，并确保它能够合理解释数字生物标记与数字表型之间的关系。常用的验证方法如下[22]：

（1）因子分析（factor analysis，FA）。因子分析用于检测多个观察变量是否可以归因于更少的潜变量。其模型为 $X=L\boldsymbol{F}+\dot{o}$，其中，$X\in\mathbf{R}^{N\times d}$ 是 N 个样本的 d 维可观测变量，$L\in\mathbf{R}^{N\times k}$ 是 k 维潜变量（$k<d$），$\boldsymbol{F}\in\mathbf{R}^{k\times d}$ 是因子载荷矩阵，表示潜变量对可测变量的影响，$\dot{o}$ 是误差。采用 MLE 计算因子载荷 $L^*=\mathrm{argmax}_L P(X|L)$，采用 Bartlett 因子得分法估算潜变量值 $L=(F^{\mathrm{T}}\Sigma^{-1}F)^{-1}F^{\mathrm{T}}\Sigma^{-1}X$。通过 Kaiser-Meyer-Olkin（KMO）检验潜变量适用性：KMO>0.7 表示潜变量适用；通过 Bartlett 球形检验相关结构，拒绝原假设（$P<0.05$）表示变量之间存在相关结构。

（2）结构方程模型。结构方程模型（structural equation modeling，SEM）用于建立可测变量 X 和潜变量 L 之间的因果关系：$Y=\beta L+\gamma X+\dot{o}$，其中 Y 是目标变量（如疾病状态），β、γ 是路径系数。使用广义最小二乘法（generalized least squares，GLS）或 MLE 估计路径系数：$\hat{\theta}=\mathrm{argmin}_\theta(Y-\beta L-\gamma X)^{\mathrm{T}}W(Y-\beta L-\gamma X)$。通过拟合优度（goodness-of-fit，GOF）评价模型优度：RMSEA<0.05（较好），0.05≤RMSEA<0.08（可接受）；CFI/NFI>0.9 表示模型拟合良好；通过 t 统计量检验路径显著性检验，$P<0.05$ 说明路径有效。

（3）潜变量回归（latent variable regression，LVR）。用于验证潜变量 L、DM 是否对 DP 具有预测价值。计算公式为：$\mathrm{DP}=\alpha L+\beta\mathrm{DM}+\gamma Z+\dot{o}$，其中 Z 是协变量（如年龄、性别），β，γ 是回归系数。使用偏最小二乘回归（partial least squares regression，PLSR）估计模型：$\hat{\beta}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$。通过模型解释度 R^2 评价解释效果，$R^2>0.5$ 表示潜变量可解释大部分数据变异，通过标准误差（standard error，SE）评价模型稳定性，标准误差较小表示模型稳健。

（4）因果推理验证。由于潜变量无法直接观测，需要因果推理评估其影响。采用 DAG 发现潜变量：$P(\mathrm{DP}|\mathrm{do}(L))$，通过贝叶斯网络学习因果结构 DAG：

$$P(\mathrm{DP}|\mathrm{do}(\mathrm{DM}))=\sum_L P(\mathrm{DP}|L)P(L|\mathrm{do}(\mathrm{DM}))。\quad(27)$$

采用 Granger 因果检验[35]评估 L 是否是潜变量。通过反事实分析[17]，计算个体层面潜变量的影响：$\mathrm{DP}_{L=1}-\mathrm{DP}_{L=0}$，如果 $\Delta\mathrm{DP}>0$，则 L 确实影响 DP；如果 $\Delta\mathrm{DP}\approx0$，则 L 可能不是关键潜变量。

3.3.3.5 稳定性

稳定性是数字生物标记和数字表型需要满足的重要准则之一。它要求在不同环境、个体、测量设备或数据采集条件下，数字生物标记和数字表型的因果效应应保持稳定，即其估计结果在内在一致性和外在稳健性之间不发生显著偏移。

在因果推理框架下，希望保证（DM→DP）关系在不同数据环境和测量条件下保持稳定。可以表示为：$P(\mathrm{DP}|\mathrm{do}(\mathrm{DM}),E_1)=P(\mathrm{DP}|\mathrm{do}(\mathrm{DM}),E_2)$；对于任意两个数据环境 E_1 和 E_2，如果条件干预效应

保持相等,则说明该因果关系在不同环境下保持稳定。定义条件因果效应(conditional average treatment effect, CATE)为 $CATE(E)=E[DP \mid do(DM),E]$,如果 $CATE(E)$ 在不同环境 E 中保持不变,则因果效应是稳定的。用稳定性偏移度评价稳定程度:$S_{bias}=|CATE(E_1)-CATE(E_2)|$,如果 $S_{bias}\approx 0$,则数字生物标记和数字表型在不同环境下稳定。

稳定性可以通过一致性分析进行评估,包括内在一致性和外在一致性评估:

(1) 内在一致性分析。内在一致性要求在相同个体或数据环境下,多次测量的因果效应应保持稳定。常用试验重复测量法:定义同一患者在不同时间点 t_1 和 t_2 下测量的数字生物标记:$DM_{t_1}=DM_{t_2}+\grave{o}$,如果 $\grave{o}\sim N(0,\sigma^2)$ 具有零均值正态分布,则该生物标记具有稳定性。对于不同个体间的稳定性,可以计算类间相关系数(intraclass correlation coefficient, ICC):$ICC=\frac{\sigma_B^2}{(\sigma_B^2+\sigma_W^2)}$,其中,$\sigma_B^2$ 是不同个体间的方差;σ_W^2 是相同个体内的测量方差。如果 $ICC>0.8$,则说明数字生物标记具有高一致性和稳定性。

(2) 外在一致性分析。外在一致性要求在不同数据环境(如设备、测量方法、时间、地点)下,数字生物标记和数字表型的因果效应应保持稳定。包括两类检验方法:① 多环境因果推理(invariant causal prediction, ICP)。用于检查因果关系在不同环境 E 下的稳定性:$P(DP \mid do(DM),E_1)=P(DP \mid do(DM),E_2)$;具体方法可采用采用因果结构学习方法(如 DAG 推理)识别因果稳定性:$G(E_1)=G(E_2)$,其中,$G(E)$ 代表因果图结构;使用条件独立性(conditional independence, CI)测试验证因果效应是否稳定:$P(DP \mid DM,E)\perp E$,若该条件成立,则 DM→DP 因果关系在不同环境 E 下不受影响,说明其因果效应稳定。②Bland-Altman 统计一致性分析。用于评估两种测量方法或环境下的稳定性:$LOA=\bar{d}\pm 1.96\cdot SD_d$,其中,$d=X_1-X_2$ 代表不同测量条件下的数值差异,$\bar{d}$ 是均值偏差,SD_d 是标准差;如果 LOA 在可接受范围内(如小于 5%变异),则说明该数字生物标记或数字表型在不同环境下具有稳定性。

为了确保因果效应的稳健性,通常还需进一步进行稳健性检验。包括:

(1) 多环境稳健性估计。用于评估因果效应在不同环境下的稳定性,表述为 $\max_{E_1,E_2}|P(DP \mid do(DM),E_1)-P(DP \mid do(DM),E_2)|<\grave{o}$。这里 $\grave{o}$ 是可接受的稳定性偏差阈值(如 5%);如果该公式成立,则说明因果效应在不同环境下是稳健的。

(2) 反事实稳健性估计。$E[DP_{DM=1}-DP_{DM=0} \mid E_1]=E[DP_{DM=1}-DP_{DM=0} \mid E_2]$,由此计算不同环境下的因果效应均值,如果它们相等,则说明因果效应在不同环境下稳定。

3.3.3.6 终点性:因果推理预测疾病替代终点

数字表型作为临床替代终点的终点性准则验证方法,包括 Prentice 准则和因果中介分析,并结合 SCM 和贝叶斯网络进行验证。

(1) Prentice 准则。Prentice[23](1989)提出了一套判定某变量 S(这里指数字表型)是否可以作为替代终点的准则:若(替代终点)能够完全解释治疗/干预 X 对临床主要终点 T 的影响,则 S 是合适的替代终点。此时:$P(T \mid X,S)=P(T \mid S)$,即在控制 S 之后,X 对 T 没有额外效应。Prentice 准则的四个条件:① 干预 X 影响替代终点 S,即 $X\rightarrow S$, $P(S \mid X)\neq P(S)$;② 干预 X 影响主要终点 T,即 $X\rightarrow T$, $P(T \mid X)\neq P(T)$;③ 替代终点 S 影响主要终点 T,即 $S\rightarrow T$, $P(T \mid S)\neq P(T)$;④ 控制 S 后,X 对 T 无额外效应,即 $P(T \mid X,S)=P(T \mid S)$。如果同时满足所有四个条件,则 S 是 T 的合适替代终点。

(2) 因果中介分析[24]可以进一步量化替代终点 S 是否合理。其核心是分解总因果效应(total effect, TE)为 NIE 和 NDE。假设 X 为干预(如治疗措施)、S 为替代终点(这里指数字表型)、T 为主要临床终点、U 为潜在混杂因素,则因果模型:$T=\alpha_0+\alpha_1X+\alpha_2S+\varepsilon_T$,$S=\beta_0+\beta_1X+\varepsilon_S$。此时,TE 为 $TE=E[T \mid X=1]-E[T \mid X=0]=\alpha_1+\alpha_2\beta_1$,即干预 X 通过直接和间接途径对主要终点 T 的影响;NIE 为 $NIE=\alpha_2\beta_1$,即 X 通过 S 影响 T 的效应大小;NDE 为 $NDE=\alpha_1$,即 X 通过其他路径(不经过 S)对 T 产生的影响。如果 NIE 在总效应中占主要部分,则 S 是有效的替代终点:

$$\frac{NIE}{TE}=\frac{\alpha_2\beta_1}{(\alpha_1+\alpha_2\beta_1)}\approx 1。 \quad (28)$$

若 $NDE=\alpha_1$ 仍然具有统计学意义,则说明 S 不能完全代理 X 对 T 的效应,数字表型的终点性准则不能得到验证。

(3) SCM 分析[17]。可以使用 DAG 表示因果关系:$X\rightarrow S\rightarrow T$, $X\rightarrow T$;其中,如果 X 对 T 的效应完全通过 S 传递(即 $X\rightarrow S\rightarrow T$,则 S 是理想的替代终点。此时,可以通过计算反事实均值差检测数字表型的终点性准则:$ATE=E[T \mid do(X=1)]-E[T \mid do(X=0)]$;根据因果推断的 do-算子,如果 $E[T \mid do$

$(X=1)]-E[T|\mathrm{do}(X=0)]\approx E[T|\mathrm{do}(S=1)]-E[T|\mathrm{do}(S=0)]$，则 S 是 T 的有效替代终点。

（4）贝叶斯因果网络（Bayesian causal network，BCN）。可通过构建 BCN[36]，用贝叶斯更新法进一步计算因果路径上的信息传递情况：$P(T|X)=\sum_S P(T|S)P(S|X)$，如果 $P(T|X)$ 主要由 $P(S|X)$ 和 $P(T|S)$ 贡献，则 S 具有替代终点的价值。这可以通过计算信息熵 $H(T|X)-H(T|S)=I(T;S|X)$ 进一步定量化评估，若互信息 $I(T;S|X)$ 较大，则 S 是 T 的较好替代终点。

（5）基于真实世界数据（real-world data，RWD）的机器学习验证。多模态数据中，可采用以下方法进一步验证：①用贝叶斯信息准则（Bayesian information criterion，BIC）计算替代终点的解释力，$BIC=-2\log L+k\log N$，若引入 S 后 BIC 显著降低，说明 S 对 T 预测有贡献。②通过 Shapley 值，计算 S 在机器学习预测 T 中的重要性：

$$\varphi_i=\sum_{S\subseteq N\setminus\{i\}}\frac{|S|!\ (|N|-|S|-1)!}{|N|!}(V(S\cup\{i\})-V(S)),\tag{29}$$

若 S 贡献较大，则说明其能作为有效的替代终点。

若上述所有方法均表明 S 具有良好的终点性，则可将其作为临床替代终点应用于精准医疗和智能医学等场景。

4 案例：传染病主动监测 AI 大模型

4.1 背景意义

如图 5 所示，猩红热在人群中隐匿传播，难以及时发现。流行初期，早期患者在基层医院就诊时，因医生诊断水平有限，常被漏诊，未能及时进入国家传染病报告监测系统，导致漏诊病例返回人群，引发新感染，形成恶性循环，加剧流行扩散。因此，在流行初期依托基层门诊病历的多模态数据，构建 AI 语言表征的向量化队列数据库，基于向量化数字组学队列并借助开源大模型（如 DeepSeek、千问、豆包），通过低秩适配（low-rank adaptation，LoRA）微调，构建猩红热主动监测 AI 大模型，具有重要流行病学和公共卫生实践意义。模型通过自然语言处理（natural language processing，NLP）技术提取病历文本信息，融合实验室检查、影像报告等，建立多模态向量化数据库，并采用 Transformer 架构进行特征融合。AI 大模型实时分析门诊病历，识别疑似病例并对接传染病报告系统，提高早期发现率，减少漏诊，实现早发现、早报告、早干预，降低流行风险。

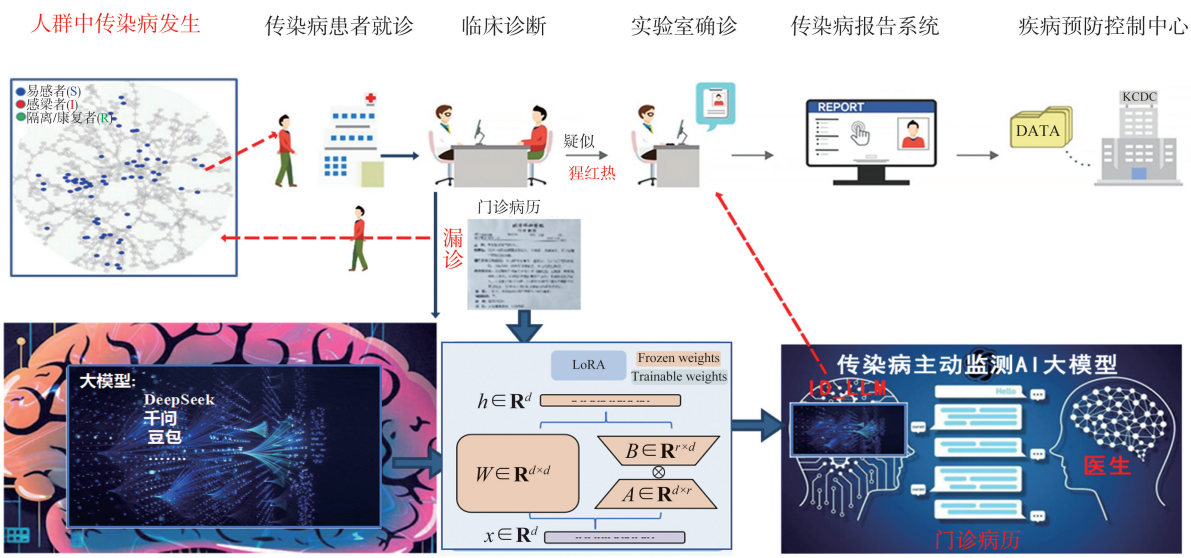


图 5 猩红热主动监测 AI 大模型的背景意义及实现途径

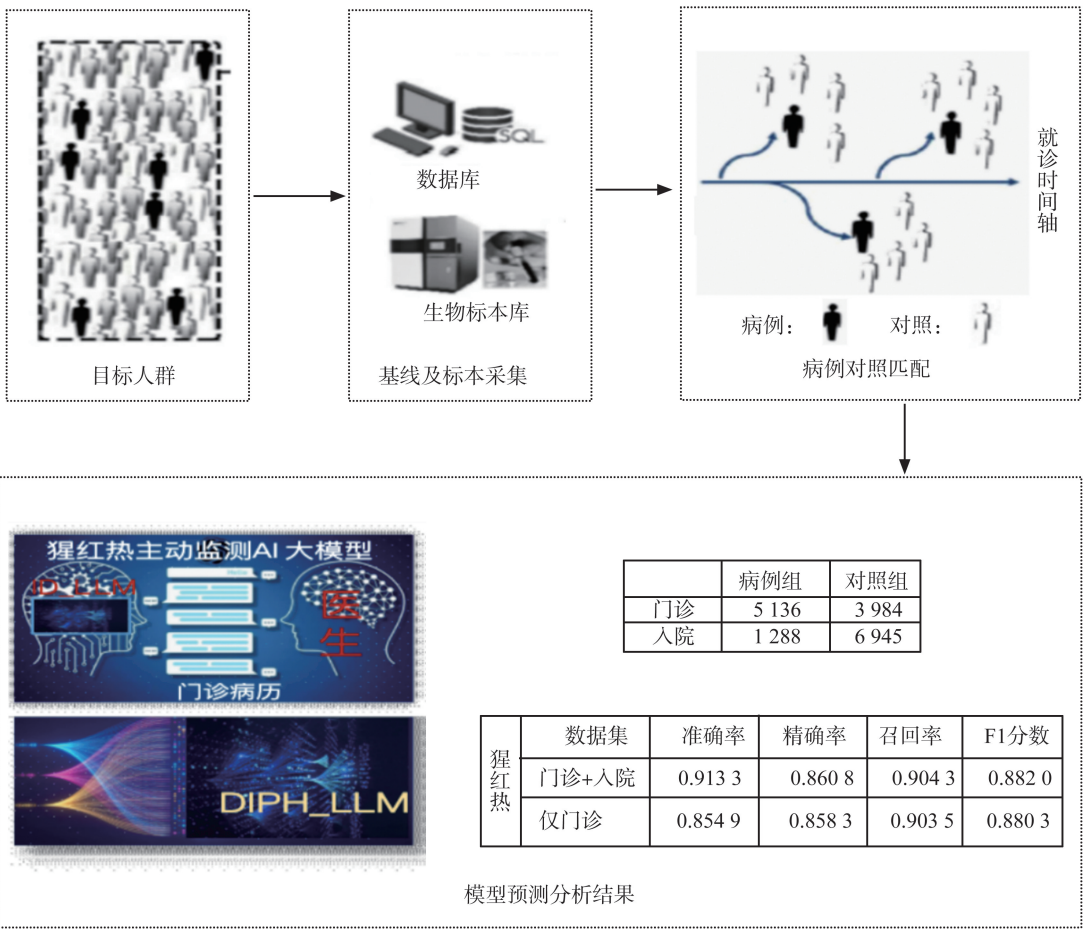
Figure 5 Background significance and implementation pathways of AI large language model for scarlet sever active monitoring

4.2 设计原理与方法

4.2.1 多模态队列创建

依托国家健康医疗大数据中心（北方）的联邦多中心健康医疗大数据协作网络，选取山东省多家不同级别医院的电子病历数据。在就诊信息表中，筛选门诊和住院记录，提取病史、主诉等核心信息，拼接形成完整的病例描述，以构建高质量的非结构化电子病历数据集。在化验信息表中，提取化验和检查结果，并与临床专家协作筛选关键诊断指标，经过数据清理和标准化处理，最终形成

结构化化验数据。基于此，建立一个多中心、多模态猩红热数据库，涵盖结构化和非结构化数据，全面整合猩红热相关的多模态信息。如图 6 所示，该数据库采用巢式病例对照研究设计，其核心思想是沿着患者的就诊时间轴，按照 1∶1至 1∶4（即 1 个病例匹配 1～4 个对照）的方式，匹配患者的就诊时间、就诊科室、年龄和性别，并通过数据链接与映射，构建统一的多模态数据库，以支持猩红热诊断特征挖掘与智能分析。猩红热病例组 6 424 例，对照组 10 929 例，总计 17 353 例。



	病例组	对照组
门诊	5 136	3 984
入院	1 288	6 945

	数据集	准确率	精确率	召回率	F1分数
猩红热	门诊+入院	0.913 3	0.860 8	0.904 3	0.882 0
	仅门诊	0.854 9	0.858 3	0.903 5	0.880 3

图 6　基于巢式病例对照设计的猩红热多模态数据链接映射及猩红热主动监测 AI 大模型的评价结果

Figure 6　Multimodal data linkage mapping for scarlet fever based on nested case-control design and evaluation results of AI large language model for scarlet fever active monitoring

4.2.2　模型构建

如图 7 所示，猩红热主动监测 AI 大模型的建模原理与方法，包括“多模态数据向量化、预训练模型、LoRA 微调及传染病主动监测推理”四大关键环节。

（1）多模态数据向量化队列数据库构建。猩红热监测数据来源于电子病历，包括门诊病历、实验室检查结果、影像数据等。这些数据具有多模态、非结构化、异构性强的特点，因此需要通过以下步骤进行向量化处理：① 采用 Transformer 架构，通过自注意力机制提取文本、影像和检验指标的语义特征。② 词向量、句向量及数值指标被映射至统一的高维向量空间，形成猩红热多模态向量化队列数据库。

（2）预训练模型。基于大规模的猩红热相关数据，利用现有的开源大模型（DeepSeek、千问、豆包）进行预训练：① 通过大规模无监督学习，让模型在电子病历、临床指南、流行病学数据等文本数据上进行深度学习，提高对传染病相关知识的理解能力。② 预训练模型能够学习到医学领域的专业表述，使其在后续微调时具备更强的泛化能力。

（3）LoRA 微调。为了提高猩红热主动监测模型的适应性，同时降低计算成本，基于猩红热多模态向量化队列数据库所提供的信息，采用 LoRA 技术对预训练大模型进行微调：① LoRA 微调通过引入额外的低秩矩阵调整模型权重，而无需更新整个神经网络，从而在计算资源受限的情况下，提高模型针对猩红热分类、预警和监测的能力。② 在此过程中，AI 模型需区分感染与非感染病例，即基于训练数据完成“猩红热 *vs.* 非猩红热”的推理优化。

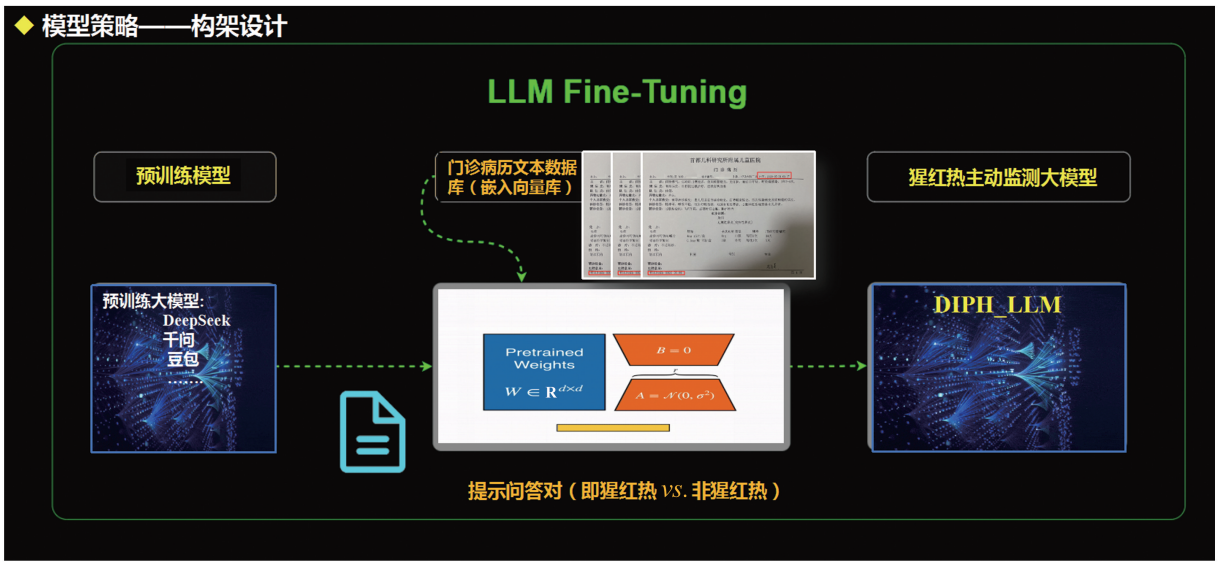


图 7 猩红热主动监测 AI 大模型的建模原理与方法

Figure 7 Modeling principles and methods of AI large language model for scarlet fever active monitoring

(4)猩红热主动监测推理。最终,通过猩红热多模态队列引导的 LoRA 微调 AI 大模型,实现猩红热主动监测:① AI 系统能够对新发病例进行快速分类和风险评估,并结合电子病历和实验室数据,生成早期预警信号。② 该系统可集成至医疗机构的传染病监测系统,辅助医生提高识别传染病的准确性,减少漏诊,提升猩红热监测能力。

4.3 结果及评价

猩红热主动监测 AI 大模型基于多家医院的多模态数据进行外部交叉验证,训练集包含门诊病历+入院记录,验证集仅使用门诊病历,以评估模型的泛化能力。模型利用 LoRA 微调,结合多模态病历数据进行猩红热识别,并通过外部交叉验证评估性能。

4.3.1 模型评价

模型采用准确率、精确率、召回率和 F1 分数作为评价指标,并针对 6 种传染病在不同数据集(门诊+入院 *vs.* 仅门诊)上的结果进行比较。结果发现,针对猩红热的训练数据:准确率 = 0.913 3,F1 分数 = 0.882 0;验证数据:准确率 = 0.854 9,F1 分数 = 0.880 3。由此可见,召回率达到 0.904 3,表明 AI 模型在监测猩红热方面具有较强的识别能力。

4.3.2 模型优化方向

(1) 针对召回率低的疾病,应补充额外的实验室检查数据,提高识别能力。

(2) 采用动态调整匹配策略,增强对不同传染病的适应性,例如利用多轮微调,使 AI 逐步学习不同疾病的特征。

4.3.3 结论

猩红热主动监测 AI 大模型经过多家家医院的外部交叉验证,在大部分疾病上表现优异,F1 分数维持在 0.88 左右,体现了 AI 在猩红热早期筛查中的巨大潜力。结果表明,仅基于门诊病历即可获得较好的监测效果,可大幅提高猩红热预警能力,同时减少住院数据依赖,提升模型的可推广性。这一研究对猩红热早期预警和防控具有重要应用价值。

5 展望与挑战

AI 语言表征的多模态队列的核心,是以统一的语言建模方式对图像、文本、音频、基因、传感器等多种模态信息进行结构化表征与高效压缩。特别是在高维嵌入空间中实现无损压缩,是实现大规模多模态分析与共享的关键路径。

未来发展的趋势,是以预训练语言模型为统一编码器,对不同模态的数据进行协同建模与压缩。近期研究已证实,Transformer 类模型对图像、文本、音频的字节序列联合建模,不仅能提取高层语义信息,还可作为统一的多模态熵估计器,实现超越传统压缩器(如 JPEG、FLAC)的跨模态通用压缩[37]。在此基础上,交叉模态预测压缩进一步释放语言模型的潜力,将文本、图像等模态作为彼此的边信息,在压缩图像时参考其描述文本,可极大减少冗余位[38]。

在嵌入层面,随着 CLIP、Flamingo、BioViL 等跨模态模型普及,嵌入向量的压缩已成为多模态队列的关键任务。一方面,采用自动编码器或聚类策略对嵌入降维并进行可学习熵编码[39],可在保证语义

保留的同时压缩数倍以上。另一方面,哈希编码技术[40]已能将嵌入向量压缩为 128 位以内的二进制码,在图文检索等任务中实现数十倍压缩比而性能不降低。此外,面向需要完全还原精度的场景,如知识图谱嵌入传播、医用语义匹配等,预测压缩方法可实现浮点向量的逐位熵编码,保障精度完整性[41]。

尽管应用前景广阔,挑战仍十分显著。①从理论层面,跨模态的联合熵建模仍缺少统一表达框架,尤其在非对称、时序性强的模态间(如语音-图像-文本)构建稳定联合分布模型仍存困难。②当前压缩模型对模态组合、应用场景依赖性强,泛化能力不足。压缩算法往往需为特定模态或任务单独训练模型,缺乏通用性。③计算资源瓶颈阻碍了语言表征压缩模型在边缘设备等低功耗场景的应用[39]。多数模型依赖大规模参数、序列预测和逐位编码,难以满足实时传输和轻量部署需求。对此,需进一步探索模型精简、推理蒸馏及边云协同架构。更为关键的是,在医学等高敏场景中,数据的合规、安全、解释性等要求对压缩模型提出新标准。模型不能仅优化码率,还需保障压缩后数据的法律可用性与伦理合规[42]。如果压缩过程导致语义丢失、诊断信息误导,将带来严重后果。

未来,AI 语言表征的多模态队列发展方向应聚焦于以下几方面:①构建统一的多模态编码语言模型,支持开放模态与异构格式的语义协同压缩;②开发任务感知压缩机制,仅保留对下游模型/推理任务有效的表征位;③推动边缘智能部署能力,设计低功耗、低延迟的轻量化压缩解压模块;④建立模型合规性与可解释性评估体系,确保 AI 模型压缩不仅能用还要更可信。

总之,AI 语言表征的多模态队列正从语言、图像、语音向结构化嵌入与联合压缩方向快速推进。借助信息论、预训练语言模型与自监督学习框架的融合,其理论基础与实际应用将迎来新突破。面向下一代 AI 系统,围绕多模态向量无损压缩的理论完备性、方法泛化性与应用安全性,将成为推动 AI 语言表征的多模态队列持续演进的关键驱动力。

参考文献:

[1] Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence[J]. Nature, 2023, 616(7956): 259-265.

[2] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[3] Huang KX, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission [EB/OL]. (2020-11-29)[2025-05-15]. https://arxiv.org/abs/1904.05342

[4] Peng YF, Yan SK, Lu ZY. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets[C]//Proceedings of the 18th BioNLP Workshop and Shared Task. Florence, Italy: Stroudsburg, PA, USAACL, 2019: 58-65.

[5] Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing [J]. ACM Trans Comput Healthcare, 2022, 3(1): 1-23.

[6] Shin HC, Zhang Y, Bakhturina E, et al. BioMegatron: larger biomedical domain language model [EB/OL]. (2020-10-14) [2025-05-15]. https://arxiv.org/abs/2010.06060

[7] Yang X, Pournejatian NM, Shin HC, et al. GatorTron: a large clinical language model to unlock patient information from unstructured electronic health records [EB/OL]. (2022-03-14) [2025-05-01]. https://arxiv.org/abs/2203.03540v2

[8] Peng C, Yang X, Chen AK, et al. A study of generative large language model for medical research and healthcare [J]. NPJ Digit Med, 2023, 6(1): 210. doi:10.1038/s41746-023-00958-w

[9] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge [J]. Nature, 2023, 620(7972): 172-180.

[10] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision [EB/OL]. (2021-02-26)[2025-05-15]. https://arxiv.org/abs/2103.00020

[11] Wang ZF, Wu ZB, Agarwal D, et al. MedCLIP: contrastive learning from unpaired medical images and text[J]. Proc Conf Empir Methods Nat Lang Process, 2022, 2022: 3876-3887. doi:10.18653/v1/2022.emnlp-main.256

[12] Feliandra ZB, Khadijah S, Rachmadi MF, et al. Classification of stroke and non-stroke patients from human body movements using smartphone videos and deep neural networks [C]//2022 International Conference on Advanced Computer Science and Information Systems (ICACSIS). Depok, Indonesia: IEEE, 2022: 187-192.

[13] Qiu ZB, Wang HX, Liao CB, et al. Sound recognition of harmful bird species related to power grid faults based on VGGish transfer learning[J]. J Electr Eng Technol, 2023, 18(3): 2447-2456.

[14] Umirzakova S, Ahmad S, Mardieva S, et al. Deep learning-driven diagnosis: a multi-task approach for segmenting stroke and Bell's palsy[J]. Pattern Recognit, 2023, 144: 109866. doi:10.1016/j.patcog.2023.109866

[15] Bannur S, Hyland S, Liu QC, et al. Learning to exploit temporal structure for biomedical vision-language processing[C]//2023 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 15016-15027.
[16] Boecking B, Usuyama N, Bannur S, et al. Making the most of text semantics to improve biomedical vision—language processing [C]//Computer Vision—ECCV 2022. Switzerland: Springer Nature, 2022: 1-21.
[17] Pearl, J. Causality: models, reasoning, and inference[M]. Cambridge, UK: Cambridge University Press, 2000.
[18] Nomura A, Takeji Y, Shimojima M, et al. Digitalomics: towards artificial intelligence/machine learning-based precision cardiovascular medicine [J]. Circ J, 2025. doi:10.1253/circj.CJ-24-0865
[19] Balasubramaniam NK, Penberthy S, Fenyo D, et al. Digitalomics-digital transformation leading to omics insights[J]. Expert Rev Proteomics, 2024, 21(9/10): 337-344.
[20] Tamura Y, Nomura A, Kagiyama N, et al. Digitalomics, digital intervention, and designing future: the next frontier in cardiology[J]. J Cardiol, 2024, 83(5): 318-322.
[21] Sameh A, Rostami M, Oussalah M, et al. Digital phenotypes and digital biomarkers for health and diseases: a systematic review of machine learning approaches utilizing passive non-invasive signals collected via wearable devices and smartphones[J]. Artif Intell Rev, 2024, 58(2): 66. doi:10.1007/s10462-024-11009-5
[22] Anderson JC, Gerbing DW. Structural equation modeling in practice: a review and recommended two-step approach[J]. Psychol Bull, 1988, 103(3): 411-423.
[23] Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria[J]. Stat Med, 1989, 8(4): 431-440.
[24] Rudolph KE, Williams NT, Diaz I. Practical causal mediation analysis: extending nonparametric estimators to accommodate multiple mediators and multiple intermediate confounders[J]. Biostatistics, 2024, 25(4): 997-1014.
[25] Alayrac JB, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning[EB/OL]. (2022-11-15)[2025-05-15]. https://arxiv.org/abs/2204.14198
[26] Yang ZC, Wei T, Liang Y, et al. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images[J]. Nat Commun, 2025, 16(1): 2366. doi:10.1038/s41467-025-57587-y
[27] Golovanevsky M, Eickhoff C, Singh R. Multimodal attention-based deep learning for Alzheimer's disease diagnosis[J]. J Am Med Inform Assoc, 2022, 29(12): 2014-2022.
[28] Wang Q, Chen K. Multi-label zero-shot human action recognition *via* joint latent ranking embedding[J]. Neural Netw, 2020, 122: 1-23. doi: 10.1016/j.neunet.2019.09.029
[29] Yang L, Xu S, Sellergren A, et al. Advancing multimodal medical capabilities of Gemini[EB/OL]. (2024-05-06)[2025-05-15]. https://arxiv.org/abs/2405.03162
[30] Oudin A, Maatoug R, Bourla A, et al. Digital phenotyping: data-driven psychiatry to redefine mental health [J]. J Med Internet Res, 2023, 25: e44502. doi: 10.2196/44502
[31] Talukder AK, Schriml L, Ghosh A, et al. Diseasomics: actionable machine interpretable disease knowledge at the point-of-care[J]. PLoS Digit Health, 2022, 1(10): e0000128. doi:10.1371/journal.pdig.0000128
[32] Molina C, Prados-Suarez B. Digital phenotypes for personalized medicine[J]. Stud Health Technol Inform, 2021, 285: 141-146. doi:10.3233/SHTI210587
[33] Myszewski JJ, Klossowski E, Meyer P, et al. Validating GAN-BioBERT: a methodology for assessing reporting trends in clinical trials[J]. Front Digit Health, 2022, 4: 878369. doi:10.3389/fdgth.2022.878369
[34] Gharavi E, LeRoy NJ, Zheng GT, et al. Joint representation learning for retrieval and annotation of genomic interval sets[J]. Bioengineering, 2024, 11(3): 263. doi:10.3390/bioengineering11030263
[35] Shojaie A, Fox EB. Granger causality: a review and recent advances[J]. Annu Rev Stat Appl, 2022, 9(1): 289-319.
[36] Zeng ZX, Jiang X, Neapolitan R. Discovering causal interactions using Bayesian network scoring and information gain[J]. BMC Bioinformatics, 2016, 17(1): 221. doi:10.1186/s12859-016-1084-8
[37] Heurtel-Depeiges D, Ruoss A, Veness J, et al. Compression via pre-trained transformers: a study on byte-level multimodal data[EB/OL]. (2024-10-07)[2025-05-15]. https://arxiv.org/abs/2410.05078
[38] Mital N, Özyilkan E, Garjani A, et al. Neural distributed image compression using common information[EB/OL]. (2021-11-10)[2025-05-15]. https://arxiv.org/abs/2106.11723
[39] Shao ZH, Wang PY, Zhu QH, et al. DeepSeekMath: pushing the limits of mathematical reasoning in open language models[EB/OL]. (2024-04-27)[2025-05-15]. https://arxiv.org/abs/2402.03300
[40] Liao SY, Chen J, Wang YZ, et al. Embedding compression with isotropic iterative quantization[J]. Proc AAAI Conf Artif Intell, 2020, 34(5): 8336-8343.
[41] Gomes C, Brunschwiler T. Neural embedding compression for efficient multi-task earth observation modelling [C]//IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium. Athens, Greece: IEEE, 2024: 8268-8273.
[42] Javed HT, Khan KU, Cheema MF, et al. Instance-based lossless summarization of knowledge graph with optimized triples and corrections (IBA-OTC)[J]. IEEE Access, 2023, 12: 5584-5604.

(编辑:相峰)