·综 述·

数据标注在公共卫生及临床医学中的应用进展

刘伊婧，张景奇

东北大学文法学院，沈阳 110169
通信作者：张景奇，Email：zhangjingqi@mail.neu.edu.cn

【摘 要】数据标注是人工智能技术应用的基础，对数据标注在公共卫生和临床医学的现实应用进行梳理，旨在为推动中国智慧医疗体系建设和公共卫生治理现代化提供参考依据。通过对中国知网（CNKI）数据库和 Web of Science 数据库核心合集系统检索，并收集 2020 年 1 月 1 日—2025 年 4 月 20 日有关数据标注在公共卫生和临床医学中应用的中英文文献。数据标注在智慧医疗管理与服务、辅助诊疗、疾病监测、药物警戒与新药研发、慢性病管理等方面展现了强大的应用潜力，能够显著提升医疗人工智能模型的准确性、优化诊疗流程以及推动精准医疗发展。当前，医疗数据标注呈现自动化、多中心协同和动态化发展趋势，但仍面临标准不统一、隐私保护等挑战。未来需构建规范化、智能化的标注生态，以充分释放人工智能技术在医疗领域潜力，助力“健康中国”战略目标的实现。

【关键词】数据标注；人工智能；公共卫生；临床医学；智慧医疗体系；数据注释

Advances in applications of data annotation in public health and clinical medicine: a review

LIU Yijing, ZHANG Jingqi (*School of Humanities and Law*, *Northeastern University*, *Shenyang* 110169, *China*)
Corresponding author: ZHANG Jingqi, E-mail: zhangjingqi@mail.neu.edu.cn

【Abstract】 Data annotation is the foundation for the application of artificial intelligence (AI). This paper reviews the real-world applications of data annotation in public health and clinical medicine, with the aim of providing reference for promoting the construction of China′s smart medical system and the modernization of public health governance. The Chinese and English articles on the applications of data annotation in public health and clinical medicine were retrieved from the China National Knowledge Infrastructure (CNKI) and the Web of Science Core Collection with the time interval from January 1, 2020 to April 20, 2025. Data annotation demonstrates significant application potential in smart healthcare management and services, diagnostic assistance, disease surveillance, pharmacovigilance and new drug development, as well as chronic disease management. It substantially enhances the accuracy of medical AI models, optimizes clinical workflows, and advances precision medicine. Currently, medical data annotation shows the development trends of automation, multi-center collaboration, and dynamization, while still facing challenges such as inconsistent standards and privacy protection. In the future, it is necessary to build a standardized and intelligent annotation ecosystem to fully release the potential of AI in the medical field and help achieve the Healthy China initiative.

【Keywords】 data labeling; artificial intelligence; public health; clinical medicine; smart healthcare system; data annotation

数据标注是一种提取和标记大数据关键信息的技术，通过对数据的筛选、清洗、分类、注释、标记和质量检验等加工处理，使其成为机器学习算法可理解的数据信息，是人工智能技术应用的基础[1]。2024 年 12 月，国家发展和改革委员会在《关于促进数据标注产业高质量发展的实施意见》[2] 中指出，“加强交通、医疗、金融、科学、制造、农业等重点行业领域数据标注，建设行业高质量数据集，支撑人工智能在行业领域的应用赋能。围绕医疗健康、人力资源、数字贸易、自动驾驶、低空经济等场景，以业务创新拉动数据标注需求。”2025 年 3 月，中国政府工作报告中强调，“深化数据资源开发利用”，而数据标注对于数据资源的深层次开发利用起到基础性支撑作用。医疗行业的数据标注是国家最为积极倡导发展的领域之一，在国家的政策导引下，数据标注已由一项行业技术逐步成长为一个专业化、智能化的朝阳产业。从产业发展现状看，中国已设立了 7 个数据标注基地（四川省成都市、辽宁省沈阳市、安徽省合肥市、湖南省长沙市、海南省海口市、河北省保定市和山西省大同市）。数据标注总规模达 17 282 TB，形成覆盖医疗、工业、教育等行业的高质量数据集 335 个，赋能 121 个国产人工智能大模型研发；引进和培育标注企业 223

开放获取：CC BY-NC-ND 4.0 DOI：10.11847/zgggws1147786
基金项目：国家社会科学基金（20CGL059）
第一作者：刘伊婧（1996—），博士在读，研究方向：公共政策。
收稿日期：2025-04-25 修回日期：2025-09-11 录用日期：2025-10-30
利益冲突：不存在 伦理审查：免除审查 出版授权：全体作者已与编辑部签署作者贡献声明及版权转让协议

家，标注从业人员达 5.8 万人，带动数据标注行业相关产值超过 83 亿元[3]。

在公共卫生及临床医学领域，大数据已深度融入公共卫生政策制定与医疗服务实践，加速推进医疗卫生服务体系优化和服务能力提升。从数据特点上看，公共卫生和临床医学数据具有高度的复杂性、专业性和私密性，需要经过前期大量的数据标注才能被人工智能所安全利用。高质量的数据标注不仅显著提升疾病诊断的准确性及效率，更能优化临床治疗方案、提高疾病预测与预防能力，并为实现精准化及个性化医疗提供重要支撑。而安全的数据标注则可以在促进政府决策者和科研工作者在探索公共卫生治理规律和优化诊疗方案的同时，避免公共卫生服务数据、患者隐私等敏感信息泄露。无论从政府公共卫生管理还是临床医学自身发展角度，数据标注都正在加速推动临床医学的智能化转型，成为开启数字健康新时代的技术基座。为更加清晰地把握数据标注技术在公共卫生和临床医学领域的最新现实应用及未来前景，研究在系统梳理数据标注概念、发展及技术流程的基础上，结合中国知网（CNKI）数据库和 Web of Science 数据库核心合集系统检索，并收集 2020 年 1 月 1 日—2025 年 4 月 20 日相关领域的研究进展，CNKI 数据库以“数据标注”“数据注释”“公共卫生”“临床医学”等为检索词，Web of Science 数据库以“data annotation”“data labling” “public health” “clinical medicine” “healthcare”等为检索词，对数据标注在智慧医疗管理与服务、辅助诊疗、疾病监控、新药开发、慢性病管理的应用进行深入分析，旨在为推动中国智慧医疗体系建设和公共卫生治理体系和治理能力现代化提供参考依据。

1 数据标注的概念和发展

1.1 *数据标注概念*　数据标注是指对原始数据进行标记、注释或分类的过程，目的是将非结构化的数据转化为结构化、机器可读的形式，以便机器学习算法能够理解和利用这些数据，可将其视为模仿人类学习过程中的经验学习，近似于人从书本中获取已有知识的认知行为[1, 4]。该过程通过人工或半自动化方式，系统性地从原始数据中提取关键特征信息，并根据具体应用场景需求赋予其相应的语义标签或注释[5]。数据标注质量与模型性能呈现显著的正相关性，高质量的标注数据能有效提升模型的预测精度，而低质量或不一致的数据则可能导致模型训练失败或性能下降。因此，数据标注作为人工智能技术落地的关键预处理环节，其质量管控在整个机器学习工作流程中具有决定性作用。

在公共卫生与临床医学领域，数据标注的核心任务是对医学影像、电子病历、基因组数据等专业医疗信息进行精准标注，如标记 CT 扫描中的肿瘤区域、分类病理切片中的细胞类型[6] 或提取临床记录中的关键诊断信息[7] 等。由于医疗数据具有高度专业性和复杂性，标注工作通常需要医生或医学专家参与，以确保标签的准确性和临床相关性。高质量的医疗数据标注能够为人工智能模型提供可靠的训练基础，使其广泛应用于疾病诊断、治疗方案推荐和医学研究等场景，相反，标注错误或不一致则可能导致严重的临床风险。

1.2 *数据标注流程*　数据标注的流程是一个系统化、多步骤的过程。以医学数据标注为例（图 1），在数据标注前，需先结合临床医学标准和具体任务需求制定数据标注规范。数据采集属于数据准备阶段[8]，采集对象通常涉及多模态医疗数据，包括医学影像、电子病历文本、病理切片、基因测序数据以及可穿戴设备采集的生理信号等。由于医疗数据的敏感性和复杂性，在标注前需进行严格的数据清洗与脱敏处理[9]。包括去除患者隐私信息、统一医学数字成像和通信（Digital Imaging and Communications in Medicine, DICOM）等医学影像格式、校正设备采集误差，以及剔除低质量或无效数据。标注环节通常由具备医学背景的专业人员完成[10]，例如涉及放射科医生标注病灶区域、病理专家标注组织类型，或临床医生标注病历中的关键信息。为提高效率，可辅以 AI 预标注工具（如肺结节自动检测），但需人工复核以确保准确性。质量审核是医疗数据标注的核心环节，通常采用三级审核机制，即初级标注人员互查、高级医师复核，以及多专家会诊争议案例。标注过程中发现的问题将反馈至标注环节修正，并同步更新标注标准。完成标注的数据将用于模型训练，并通过临床验证反馈标注质量。最后，经伦理审查和合规处理的标注数据将被安全存储，供后续科研或临床使用。

1.3 *数据标注的发展*　数据标注的发展与人工智能和机器学习的演进密切相关，经历了从手工标注到智能化标注的重要转变，并在医疗等领域展现出广泛的应用前景。数据标注的起源可追溯至 20 世纪 50 年代，随着模式识别和分类任务的兴起，研究者开始手动标注数据以训练模型。1957 年罗森布拉特提出的感知机进一步推动了其发展，但当时标注规模较小，主要用于学术研究，为后续奠定了基础。20 世纪 90 年代，随着计算机视觉和自然语言处理技术的逐渐成熟，数据标注进入快速发展阶段，众包平台通过互联网将大型标

注任务拆解并分发给大规模的线上工作者[11]，其兴起使得大规模标注成为可能。2010 年后，深度学习技术的突破推动数据标注向智能化与自动化迈进，预训练模型和主动学习等技术的应用显著提升了标注效率，同时应用场景也从计算机视觉、自然语言处理扩展至医疗[12]、金融、教育等垂直领域。

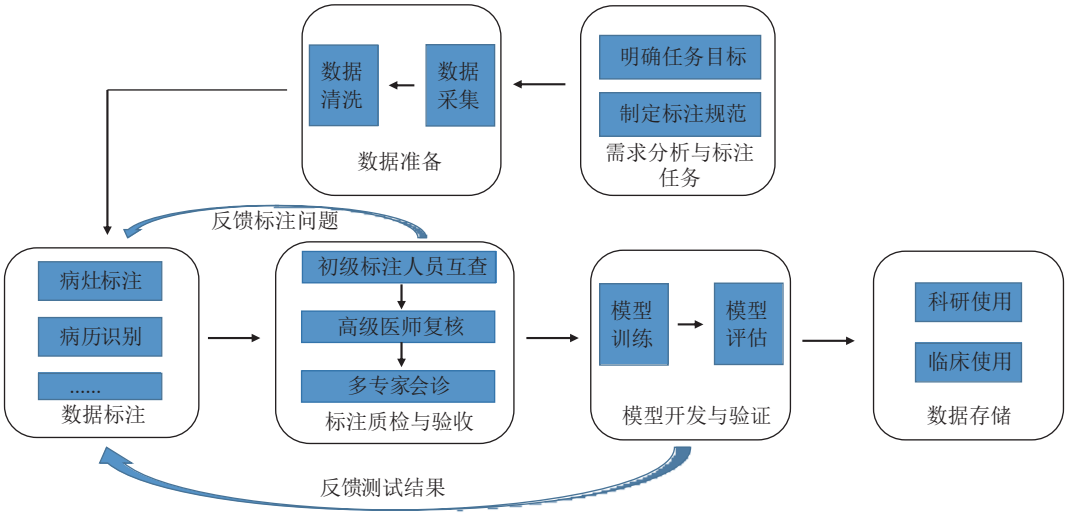


图 1 数据标注流程

Fig. 1 Data annotation process

在医疗领域，数据标注作为人工智能模型训练的基础环节，是医疗人工智能落地的核心技术支撑。随着医疗数据要素市场化及精准化需求的增长，数据标注技术在公共卫生和临床医学服务中的应用不断深化。通过精确识别和分类医疗数据，其不仅为公共卫生决策、疾病防控方案及医疗服务优化提供了有力支持，还推动了医疗智能化和个性化发展。数据标注技术的高效数据处理能力，进一步助力智慧医疗建设，持续促进人工智能技术在医疗健康领域的深度应用[13]，同时为提升政府医疗治理能力与治理体系现代化提供了重要技术支撑。从早期的小规模手工标注，到如今的智能化、自动化标注，数据标注技术的发展不仅满足了人工智能对大规模高质量数据的需求，更在医疗等关键领域推动了智能化转型，展现出广阔的应用前景。

在数据标注的助推下，国内外已经形成涵盖医学对话、医学推理、医学诊疗、中医诊断、中医药古籍、大规模多语言医疗语料库等数以万计的特色医疗数据集。社会各界对于医疗数据集的建设投入热情加大，也带动了医疗数据标注行业的快速增长。中国蚂蚁集团与上海交通大学医学院附属仁济医院泌尿外科专家团队合作，创建并正式开源中国首个医疗专科推理数据集，病种覆盖了 97.6% 以上的泌尿科就医人群，真实复刻诊疗场景。2023 年，深圳市大数据研究院发布规模最大的中文医疗问答数据集，数据集由 2 650 万个问答对组成，将现有的医疗问答数据集规模扩大了 2 个数量级，对中国生物医学问答数据集的许多现有工作进行了基准测试，并为性能最佳的模型提供开放 API。此外，中国国家基因组科学数据中心（National Genomics Data Cente, NGDC）、中国国家微生物科学数据中心（National Microbiology Data Center, NMDC）、中国国家人口健康科学数据中心（National Population Health Data Center, NPHDC）等“国家队”持续为公共卫生与医疗事业提供专业数据支撑。其中，截至 2023 年 6 月，NGDC 已存储和管理 27.6 PB 的数据资源，其开发的组学原始数据归档库（Genome Sequence Archive, GSA）被列为中国首个入选全球核心生物数据资源的数据库[14]；NMDC 牵头制定了微生物数据领域的 ISO 国际标准，并启动了“未培养微生物培养组”计划，推动了微生物数据标准的国际化进程[15]；NPHDC 则为中国首个中西医结合甲状腺专科 AI 医生“ThyroSeek 癭问”提供了核心数据资源，推动了智慧医疗落地，这些国家级数据中心的成就共同为中国公共卫生与医疗事业的高质量发展提供了坚实的数据基础和技术保障。

2 数据标注在公共卫生与临床医学领域的应用

从疾病诊断到药物研发，再到个性化医疗，由数据标注后形成的高质量数据集在智慧医疗管理服务升级、优化慢性病管理、推动机器视觉与大模型等应用于医学领域中发挥着不可或缺的作用。

2.1 升级智慧医疗管理与服务 智慧医疗管理

与服务是医疗行业数字化转型的演进方向，其本质在于通过技术赋能重构传统医疗模式，实现更高效、精准和普惠的医疗服务。面对传统医疗体系长期存在的资源分配结构性失衡、诊疗流程低效化、医患信息不对称等问题，智慧医疗依托数据驱动和技术创新，正逐步突破时空壁垒，重塑医疗资源配置范式。

数据标注作为人工智能技术落地医疗场景的关键环节，为医院智慧管理系统和诊疗全流程数字化监管提供了重要技术支撑[16]。主要体现在以下三个方面：一是优化资源配置，以西班牙加泰罗尼亚的 eConsulta 远程医疗服务为例，该系统通过标注医患对话中的症状描述、诊疗建议等关键信息，不仅构建了高效的 AI 分诊系统，显著提升了在线问诊效率，还为医疗资源配置决策提供了科学依据，是医疗数据驱动决策的典型案例[17]。二是提升管理效率，Syn-Mediverse 数据集通过合成数据技术实现了自动化、多任务标注（如物体检测、语义分割等），既克服了人工标注的成本和隐私限制，又加速了医疗机器人和手术室导航系统的研发进程，为医疗设备智能调度、手术流程优化等管理任务提供了精准的环境感知能力，从而提升医院运营效率[18]。三是改善患者体验，通过对患者问题类型（如诊断咨询、治疗方案）和医生应答策略（如专业指导、情感支持）的智能标注，有效优化医患沟通质量，实现需求精准匹配，提升服务满意度[19]。

2.2 *提升辅助诊疗能力*　临床医生与医学专家等通过对海量医疗数据进行精细化标注，构建高质量的训练数据集，从而为机器学习模型提供基础支持，实现疾病诊断、治疗方案推荐和预后预测等多种辅助诊疗功能。数据标注通过对电子病历、医学影像等数据的深度处理，识别出与疾病相关的实体，如体征、症状、检查结果、治疗效果等，进而构建出高效的疾病诊断模型。已有研究表明，基于标注的医学影像数据训练的深度学习模型，能够辅助医生快速、准确地识别疾病，显著提高肺结节[20]、乳腺癌[21] 等疾病的早期诊断率，减少误诊和漏诊的可能性。LNDb 数据库由葡萄牙研究机构和医院合作建立，数据库中的肺部 CT 图像经放射科医生和专家的专业标注，能够快速识别出结节的位置、大小、形态及恶性程度等信息，为开发肺癌早期检测和诊断工具提供了重要支持[22]。乳腺癌是女性癌症死亡的主要原因之一，国外学者利用传统自然语言处理技术与 IBM Watson 自然语言分类器，对乳腺病理报告进行自动化标注（如提取病变类型、分级等标签），构建高质量标注数据集，进而支持基于深度学习的乳腺 X 射线影像分析[23]。这种技术尤其适用于医疗资源匮乏的地区，能够显著提高乳腺癌筛查的覆盖率和准确性。在脑部影像分析中，Swinburne N C 等[24] 提出基于半监督学习的脑部 MRI 肿瘤检测模型，利用从医院影像存档与通信系统中采集的临床标注数据进行训练，选取 RetinaNet 和 Mask R-CNN 架构进行训练，并通过自标注方法（Self-labeling）自动识别未标注的肿瘤，模型性能显著提升，表明利用临床标注数据进行半监督学习能够显著提高肿瘤检测性能，并为未来的肿瘤分割、分类和治疗反应评估等下游任务奠定了基础。

在治疗方案推荐方面，数据标注通过分析电子病历和医学文献，能够根据患者的具体病情、病史和生理特征，设计出最适合的治疗方案，从而提高治疗效果并减少副作用。例如，针对胸部创伤患者的电子病历进行标注后，可以在患者入院后迅速区分严重和非严重的胸部损伤，为临床决策提供重要依据[25]。此外，通过对电子病历中的临床特征进行大规模分析，数据标注还可以辅助实现遗传性癫痫的早期诊断，帮助医生尽早采取针对性治疗，改善患者预后[26]。数据标注还可通过整合外部医学知识库（如 DBpedia）中的医学知识和电子病历中的患者数据，从而预测患者的住院率，帮助医生提前干预高风险患者，优化医疗资源的分配，同时提高诊断的准确性和效率[27]。数据标注为个性化治疗方案的制定提供了更科学的支持，推动了精准医疗的发展。

2.3 *提高疾病监测和控制能力*　数据标注技术为疾病监测、早期预警及追踪分析提供关键的数据支撑。在疾病监测环节，需整合多源异构数据，包括临床病例报告、实验室检测结果、社交媒体舆情、环境监测指标及人口流动轨迹，并通过结构化标注构建高质量训练数据集以支撑后续分析建模。例如，临床文本的实体识别技术可自动提取症状、疾病等关键信息，辅助构建实时监测系统，而个人身份信息标注与匿名化处理能严格保障数据隐私与合规性[28]，是数字化健康管理系统得以大规模应用实施的重要技术基础，其能在严格保障数据隐私与合规性的前提下，实现对高风险人群的精准追踪。

疾病的爆发往往伴随着某些异常事件的发生，例如某种疾病的病例数突然增加，或者某种病原体在特定区域的出现频率显著上升。数据标注有助于快速识别危机，制定有效的应对策略。以突发公共卫生事件为例，可对社交媒体数据、新闻报道、医疗系统数据等进行实时标注，识别关键信息，评估事件影响。例如，美国通过分析 X（原 Twitter）用户发布与新型冠状病毒（COVID-19）

相关的症状、暴露风险等信息，用于补充传统的监测系统，从而更全面地追踪疫情传播[29]；或是通过标注病例的接触史，识别疾病的传播路径、感染源以及感染者的活动轨迹，划出高风险群体或疾病聚集区域，为精准防控提供支持[30]；以及通过标注疫情前后社交媒体上关于心理健康讨论的内容，帮助政府理解全球事件对公众心理健康的影响，体现了公共健康政策需长期关注公众的心理健康需求[31]。

变体追踪是确定疾病传播路径和变异动态的关键环节。数据标注技术通过对病原体基因组的注释、结构差异分析以及耐药基因的识别与标注，为传染病防控和病原体变异监测提供了前沿技术支撑。在 COVID-19 疫情期间，通过对新冠病毒的病原体基因组数据持续标注、对比和变异追踪，能够快速识别新出现的变异株，解析其传播力和致病性的变化趋势，有助于政府及时监控病毒变异，作为早期预警系统，为公共卫生响应和疫苗研发提供科学指导[32]。另外，通过模拟免疫细胞谱系分化的分层注释策略，优化了免疫细胞及其亚型的高通量注释流程，部分克服了单细胞数据中因高度异质性导致的分类挑战，且该技术已成功应用于 COVID-19 患者数据分析，实现了对免疫细胞及其亚型的高效精准识别，揭示了单核细胞亚群比例异常升高等关键免疫特征，为深入解析疾病免疫机制提供了重要技术支撑[33]。

2.4 *推动药物警戒与新药研发* 药物警戒（pharmacovigilance）旨在检测、评估、理解并预防药品不良反应或任何与药物相关的问题[34]，包括用药错误、药物滥用、假冒伪劣商品等[35]。传统药物警戒方法依赖于病例报告和临床回顾，存在数据量有限、信息不完整等问题，而数据标注可以识别和标记出与药物相关的不良事件[36]，并将数据源扩充至社交媒体 X（原 Twitter）[37] 和论坛[38] 等，以此快速定位高风险药物或药物组合。例如，药物与食物成分的相互作用可能会改变药物的疗效，甚至导致不良反应，通过对“药物–食物”组合进行标注，可预测其成分之间的相互作用，帮助避免患者出现不良反应[39]。

传统的药物研发主要依赖高通量筛选和大量实验，不仅耗时且成本高昂，还导致严重的“Me too”现象，即多个药物针对同一靶点进行重复开发，造成资源浪费和创新不足[40]。通过数据标注技术，可以高效识别和优化候选药物，从而将资源重新分配到未满足的医疗需求领域，推动更具创新性和差异化的药物研发。数据标注技术，包括化合物结构标注、药物相似性评估、PAINS 警报检测和 Murcko 骨架生成等，为新型药物研发提供了重要的信息支持[41]，并已广泛应用于药物性肝损伤[42]、青光眼[43]、癌症[44] 等疾病的药物治疗中。在基于配体的药物设计（ligand-based drug design, LBDD）中，标注药物分子的化学结构和活性位点等信息，能在药物开发的早期阶段识别出具有风险的化合物，降低药物开发的安全风险[45]。在基于结构的药物设计（structure-based drug design, SBDD）中，标注靶点蛋白的三维结构和结合位点等信息，可直接设计出与靶标蛋白高度匹配的化合物，适用于新药研发[46]。

2.5 *优化慢性病管理* 慢性病管理涉及长期、多维度的数据积累，包括患者的生理指标（如血糖、血压、血脂）、生活方式（饮食、运动、睡眠）、用药记录以及影像学检查结果等，这些数据通常以非结构化或半结构化的形式存在于电子健康档案、可穿戴设备或居家监测设备中，使得慢性病的长期管理成为公共卫生治理的难点。在实际应用中，数据标注能够支持自然语言处理技术对包含患者病历信息、治疗记录、药物处方等电子健康记录（Electronic Health Records, EHR）进行抽取和结构化处理，辅助医生进行诊断和对病患健康情况的长期跟踪。例如，数据标注技术通过分析糖尿病患者电子健康记录中的他汀类药物使用情况，识别出未用药的五大原因（如副作用、临床惰性等）及其人群分布差异，为制定针对性的用药依从性干预措施提供了精准靶点，也为减少医疗不平等提供了循证依据[47]。

因此，医学数据标注能有效转化非结构化病历为公共卫生治理的知识基础设施，推动从被动医疗向主动健康管理的转型[48]。传统公共卫生监测的“一刀切”策略难以应对不同人群的差异化需求[49]，但随着可穿戴设备和移动健康应用的普及，实时采集个人健康数据成为可能[50]。结合数据标注技术，这些数据可用于分析用户行为模式，预测健康风险，并实施早期干预[51]，填补了传统数据库在日常活动场景下的空白，助力个性化健康管理[52]，为慢性病管理提供了动态性、预防性和个性化的新范式。在可穿戴设备生成的低信噪比数据场景下，数据标注的优化可显著提升机器学习模型的鲁棒性，使其能精准识别慢性病特征（如癫痫发作、血糖异常），并降低误报率[53]。例如，通过可穿戴压力鞋垫采集的步态数据，标注关键步态事件（如足跟触地、脚尖离地）和临床特征，为帕金森的早期诊断、病情监测和康复干预提供了量化依据，推动了智能管理系统的精准干预[54]。

3 结论和展望

国家发展和改革委员会提出，“到 2027 年，数

据标注产业专业化、智能化及科技创新能力显著提升，产业规模大幅跃升，年均复合增长率超过20%”的发展目标。而数据标注作为医疗人工智能发展的基石，正推动医疗卫生领域迈向智能化新纪元。公共卫生与临床医学领域人工智能与大模型的发展离不开高质量的数据集，而高质量的数据集的形成则离不开数据标注技术的升级。近年来，公共卫生与临床医学领域中的数据标注呈现出三大发展趋势：一是自动化标注的普及，基于 AI 的智能标注系统大幅提升了标注效率；二是多中心协同标注的推广，通过联邦学习支持的多机构协作，在保障数据隐私的前提下实现医疗数据的价值共享；三是动态标注的应用，对可穿戴设备采集的连续生理信号进行长期监测标注，推动慢性病管理从静态干预转向动态监测。

随着多模态标注技术的突破，通过精准标注医学影像中的病灶特征、电子病历中的临床实体以及基因组数据中的变异位点，显著提升了数据驱动的决策能力，为疾病诊断、流行病监测、医学研究和公共卫生管理等提供了关键支持。数据标注在医疗卫生领域的全面渗透不仅体现了科技与医学的深度融合，更预示着数字健康新时代的到来。然而，公共卫生和临床医学领域的数据标注仍面临标准不一、质量参差不齐、协同标注不畅等问题，相关部门应加速构建起规范化、智能化、协同化的医疗数据标注生态，健全数据标注标准、释放公共数据标注需求、打造高水平创新载体，为公共卫生与临床医学领域的数据标注产业发展提供更加充足的要素保障，充分释放人工智能等新技术在提升诊疗精度、优化医疗资源配置方面的巨大潜力，通过提升标注公共服务能力助力“健康中国”战略目标的实现。

数据可用性声明　本研究未生成新的数据。

参考文献

[1] Zhu J R, Kaplan R, Johnson J, et al. HiDDeN: Hiding data with deep networks[C]//The Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 682-697.

[2] 中华人民共和国国家发展和改革委员会. 关于促进数据标注产业高质量发展的实施意见 (发改数据〔2024〕1822 号)[EB/OL]. (2024-12-19)[2025-03-10]. https://www.ndrc.gov.cn/xxgk/zcfb/tz/202501/t20250113_1395643.html.

[3] 刘坤. 我国 7 个数据标注基地标注规模再创新高 [N]. 光明日报, 2025-03-22(03).

[4] 蔡莉, 王淑婷, 刘俊晖, 等. 数据标注研究综述 [J]. 软件学报, 2020, 31(2): 302-320.

[5] Foumani S N M, Nickabadi A. A probabilistic topic model using deep visual word representation for simultaneous image classification and annotation[J]. Journal of Visual Communication and Image Representation, 2019, 59: 195-203.

[6] Yan Z L, Huang H S, Geng R M, et al. Improving lung cancer pathological hyperspectral diagnosis through cell-level annotation refinement[J]. Scientific Reports, 2025, 15(1): 8086.

[7] Liukas T, Rosio R, Peltonen L M. Towards automated risk prediction of persistent pain: exploring psychosocial factors from electronic health records after breast cancer surgery[J]. Nursing Open, 2023, 10(5): 3399-3405.

[8] Berriel R F, Rossi F S, De Souza A F, et al. Automatic large-scale data acquisition via crowdsourcing for crosswalk classification: a deep learning approach[J]. Computers and Graphics, 2017, 68: 32-42.

[9] Boselli R, Cesarini M, Mercorio F, et al. An AI planning system for data cleaning[C]//The Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Skopje: Springer, 2017: 349-353.

[10] Jaiswal A, Katz A, Nesca M, et al. Identifying risk factors associated with lower back pain in electronic medical record free text: deep learning approach using clinical note annotations[J]. JMIR Medical Informatics, 2023, 11: e45105.

[11] 范红霞, 俞律弘. 众包平台用户数字标注行为中的偏见渗透 [J]. 新闻记者, 2024(11): 101-112.

[12] Swain S, Bhushan B, Dhiman G, et al. Appositeness of optimized and reliable machine learning for healthcare: a survey[J]. Archives of Computational Methods in Engineering, 2022, 29(6): 3981-4003.

[13] Colicchio T K, Dissanayake P I, Cimino J J. Formal representation of patients' care context data: the path to improving the electronic health record[J]. Journal of the American Medical Informatics Association, 2020, 27(11): 1648-1657.

[14] 陈梅丽, 王彦青, 李茹姣, 等. 国家基因组科学数据中心: 现状及展望 [J]. 中国科学: 生命科学, 2024, 54(6): 1088-1100.

[15] 范国梅, 孙清岚, 张荐辕, 等. 国家微生物科学数据中心——国家微生物资源大数据体系建设 [J]. 生命科学, 2023, 35(12): 1599-1607.

[16] De A, Huang M, Feng T H, et al. Analyzing patient secure messages using a fast health care interoperability resources (FIHR)-based data model: development and topic modeling study[J]. Journal of Medical Internet Research, 2021, 23(7): e26770.

[17] López Seguí F, Walsh S, Solans O, et al. Teleconsultation between patients and health care professionals in the Catalan primary care service: message annotation analysis in a retrospective cross-sectional study[J]. Journal of Medical Internet Research, 2020, 22(9): e19149.

[18] Mohan R, Arce J, Mokhtar S, et al. Syn-mediverse: a multimodal synthetic dataset for intelligent scene understanding of healthcare facilities[J]. IEEE Robotics and Automation Letters, 2024, 9(8): 7094-7101.

[19] Alhuzali H, Alasmari A, Alsaleh H. MentalQA: an annotated Arabic corpus for questions and answers of mental healthcare[J]. IEEE Access, 2024, 12: 101155-101165.

[20] Fedorov A, Hancock M, Clunie D, et al. DICOM re-encoding of volumetrically annotated Lung Imaging Database Consortium (LIDC) nodules[J]. Medical Physics, 2020, 47(11): 5953-5965.

[21] Park G E, Kim S H, Nam Y, et al. 3D breast cancer segmentation in DCE-MRI using deep learning with weak annotation[J]. Journal of Magnetic Resonance Imaging, 2024, 59(6): 2252-2262.

[22] Ferreira C A, Sousa C, Marques I D, et al. LNDb v4: pulmonary nodule annotation from medical reports[J]. Scientific Data, 2024, 11(1): 512.

[23] Trivedi H M, Panahiazar M, Liang A, et al. Large scale semi-automated labeling of routine free-text clinical records for deep learning[J]. Journal of Digital Imaging, 2019, 32(1): 30-37.

[24] Swinburne N C, Yadav V, Kim J, et al. Semisupervised training of a brain MRI tumor detection model using mined annotations[J]. Radiology, 2022, 303(1): 80-89.

[25] Kulshrestha S, Dligach D, Joyce C, et al. Prediction of severe chest injury using natural language processing from the electronic health

record[J]. Injury, 2021, 52(2): 205-212.

[26] Galer P D, Parthasarathy S, Xian J L, et al. Clinical signatures of genetic epilepsies precede diagnosis in electronic medical records of 32, 000 individuals[J]. Genetics in Medicine, 2024, 26(11): 101211.

[27] Gazzotti R, Faron-Zucker C, Gandon F, et al. Injecting domain knowledge in electronic medical records to improve hospitalization prediction[C]//The Proceedings of the 16th International Conference on Semantic Web. Portorož: Springer, 2019: 116-130.

[28] Dunstan J, Vakili T, Miranda L, et al. A pseudonymized corpus of occupational health narratives for clinical entity recognition in Spanish[J]. BMC Medical Informatics and Decision Making, 2024, 24(1): 204.

[29] Klein A Z, Magge A, O'Connor K, et al. Toward using twitter for tracking COVID-19: a natural language processing pipeline and exploratory data set[J]. Journal of Medical Internet Research, 2021, 23(1): e25314.

[30] Wang Z Z, Liu X F, Du Z W, et al. Epidemiologic information discovery from open-access COVID-19 case reports via pretrained language model[J]. iScience, 2022, 25(10): 105079.

[31] Rani S, Ahmed K, Subramani S. From posts to knowledge: annotating a pandemic-era Reddit dataset to navigate mental health narratives[J]. Applied Sciences, 2024, 14(4): 1547.

[32] Cheng Y X, Ji C Y, Zhou H Y, et al. Web resources for SARS-CoV-2 genomic database, annotation, analysis and variant tracking[J]. Viruses, 2023, 15(5): 1158.

[33] Jiang Y, Chen Z Y, Han N, et al. sc-ImmuCC: hierarchical annotation for immune cell types in single-cell RNA-seq[J]. Frontiers in Immunology, 2023, 14: 1223471.

[34] 刘巍, 陈易新. 药物警戒的概念与起源 [J]. 中国执业药师, 2008, 5(7): 16-18.

[35] 曾繁典. 世界卫生组织药物警戒指标及其应用 [J]. 医药导报, 2016, 35(11): 1159-1163.

[36] Narayanan S, Mannam K, Achan P, et al. A contextual multi-task neural approach to medication and adverse events identification from clinical text[J]. Journal of Biomedical Informatics, 2022, 125: 103960.

[37] Tekumalla R, Banda J M. Using weak supervision to generate training datasets from social media data: a proof of concept to identify drug mentions[J]. Neural Computing and Applications, 2023, 35(25): 18161-18169.

[38] Karapetiantz P, Audeh B, Redjdal A, et al. Monitoring adverse drug events in web forums: evaluation of a pipeline and use case study[J]. Journal of Medical Internet Research, 2024, 26: e46176.

[39] Kha Q H, Le V H, Hung T N K, et al. Development and validation of an explainable machine learning-based prediction model for drug-food interactions from chemical structures[J]. Sensors, 2023, 23(8): 3962.

[40] Repetto M, Crimini E, Belli C, et al. A demand-offer critical analysis of current drug development. Phase I drugs versus TCGA sequencing data[J]. European Journal of Cancer, 2023, 190: 112958.

[41] Gadiya Y, Shetty S, Hofmann-Apitius M, et al. Exploring SureChEMBL from a drug discovery perspective[J]. Scientific Data, 2024, 11(1): 507.

[42] Martin M T, Koza-Taylor P, Di L, et al. Early drug-induced liver injury risk screening: "free," as good as it gets[J]. Toxicological Sciences, 2022, 188(2): 208-218.

[43] Chen J S, Lin W C, Yang S, et al. Development of an open-source annotated glaucoma medication dataset from clinical notes in the electronic health record[J]. Translational Vision Science and Technology, 2022, 11(11): 20.

[44] Mitsopoulos C, Di Micco P, Fernandez E V, et al. CanSAR: update to the cancer translational research and drug discovery knowledgebase[J]. Nucleic Acids Research, 2021, 49(D1): D1074-D1082.

[45] Moein M, Heinonen M, Mesens N, et al. Chemistry-based modeling on phenotype-based drug-induced liver injury annotation: from public to proprietary data[J]. Chemical Research in Toxicology, 2023, 36(8): 1238-1247.

[46] Martin H J, Melo-Filho C C, Korn D, et al. Small molecule antiviral compound collection (SMACC): a comprehensive, highly curated database to support the discovery of broad-spectrum antiviral drug molecules[J]. Antiviral Research, 2023, 217: 105620.

[47] Sarraju A, Zammit A, Ngo S, et al. Identifying reasons for statin nonuse in patients with diabetes using deep learning of electronic health records[J]. Journal of the American Heart Association, 2023, 12(7): e028120.

[48] Chang H Y, Zan H Y, Zhang S, et al. Construction of cardiovascular information extraction corpus based on electronic medical records[J]. Mathematical Biosciences and Engineering, 2023, 20(7): 13379-13397.

[49] Abad Z S H, Butler G P, Thompson W, et al. Physical activity, sedentary behavior, and sleep on twitter: multicountry and fully labeled public data set for digital public health surveillance research[J]. JMIR Public Health and Surveillance, 2022, 8(2): e32355.

[50] Kishore N, Kiang M V, Engø-Monsen K, et al. Measuring mobility to monitor travel and physical distancing interventions: a common framework for mobile phone data analysis[J]. The Lancet Digital Health, 2020, 2(11): e622-e628.

[51] Lee J S, Browning E, Hokayem J, et al. Smartphone and wearable device-based digital phenotyping to understand substance use and its syndemics[J]. Journal of Medical Toxicology, 2024, 20(2): 205-214.

[52] Shen Q, Gao H X, Li Y W, et al. An open-access arrhythmia database of wearable electrocardiogram[J]. Journal of Medical and Biological Engineering, 2020, 40(4): 564-574.

[53] Zhang J W, Chatzichristos C, Vandecasteele K, et al. Automatic annotation correction for wearable EEG based epileptic seizure detection[J]. Journal of Neural Engineering, 2022, 19(1): 016038.

[54] Chatzaki C, Skaramagkas V, Tachos N, et al. The smart-insole dataset: gait analysis using wearable sensors with a focus on elderly and Parkinson′s patients[J]. Sensors, 2021, 21(8): 2821.