doi:10.3772/j.issn.2095-915x.2025.03.004

基于循证视角的高质量临床研究论文自动识别方法的进展与启示

开放科学（资源服务）标识码（OSID）

胥美美

中国医学科学院医学信息研究所 北京 100020

摘要：[目的 / 意义] 生物医学文献的快速增长为临床医生跟踪最新知识、检索高质量证据带来挑战。采用人工智能技术识别高质量临床研究，将显著促进有效证据的获取与更新。[方法 / 过程] 首先界定了临床研究及其论文质量的相关概念，随后围绕高质量临床研究论文的识别技术体系，从金标准、具体识别技术等方面对相关研究成果进行了系统梳理，最后总结当前研究现状，指出存在的问题，并对未来发展提出相应启示。[结果 / 结论] 建议未来进一步开发覆盖外科领域和最新临床研究的金标准数据集，拓展构建非英语语种高质量临床研究的自动识别模型，并完善识别高质量临床研究的人工智能方法。

关键词：临床研究；高质量研究；自动识别方法；临床证据；机器学习

中图分类号：G35；G254；G251；G203

Progress and Insights on Auto-Identifying High-Quality Clinical Research Papers from an Evidence-Based Perspective

XU Meimei

Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

Abstract: [Objective/Significance] The rapid growth of biomedical literature brings challenges to clinicians in tracking the latest knowledge and retrieving high-quality evidence. The use of artificial intelligence technology to identify high-quality clinical research will greatly improve the acquisition and update of effective evidence. [Methods/Processes] The article firstly defines the concepts of clinical research and clinical research paper quality, then sorts out relevant research results around the identification technology system of high-quality clinical research papers, from gold standards to specific identification technologies, and finally summarizes the current research status and proposes existing problems and related enlightenment for future development. [Results /Conclusions] It is recommended to develop large-scale gold standard data covering surgical fields and the latest clinical research

基金项目 中国医学科学院医学与健康科技创新工程项目“生物医学文献信息保障与集成服务平台”（2021-I2M-1-033）；四大慢病重大专项“慢病防治科技成果登记与评价信息集成平台研发（2023ZD0509705）。

作者简介 胥美美 (1987-)，博士，副研究员，主要研究方向为科技评价、循证决策，E-mail：xumeimei2005@163.com。

引用格式 胥美美 . 基于循证视角的高质量临床研究论文自动识别方法的进展与启示 [J]. 情报工程 , 2025, 11(3): 41-57.

doi:10.3772/j.issn.2095-915x.2025.03.004

in the future, expand the auto-identification model for high-quality clinical research in non-English languages, and improve the artificial intelligence method for auto-identifying high-quality clinical research.

Keywords: Clinical Research; High-Quality Research; Auto-Identification Method; Clinical Evidence; Machine Learning

引言

根据美国国家医学图书馆统计报告，2023年 PubMed 新索引的引文超过 150 万篇 [1]。生物医学文献的庞大体量加之其每年以百万余篇的速度增长，为循证医学实践带来了巨大挑战，使得医学研究人员难以跟踪各疾病领域的新知识，并且需要耗费大量时间和精力评估新近发表研究的质量，以捕捉有意义的临床证据。随着数据驱动方法的出现和人工智能技术的发展，机器学习算法开始被应用于评估生物医学文献的质量 [2-3]，特别是在辅助研究人员识别高质量临床研究方面，这将显著提升有效证据的获取与更新效率。

现有研究对论文的质量评价主要集中在论文的影响力和创新力两方面。论文影响力评价包括学术影响力和社会影响力评价 [4]，学术影响力多基于引文指标进行评价，而社会影响力主要依赖于 Altmetrics 指标来评价。论文创新力评价方法一般分为定性评价和定量评价。定性评价以同行评议为代表性方法来评价论文的创新性，也是使用最早、当前最为常用的评价方法。但同行评议方法由于对专家的依赖性较高，因此会存在主观性和不全面性等问题 [5]，为此，学者积极探索定量评价方法。定量评价方法主要包括基于引文指标和基于论文内容指标的两种方法 [6]。曹慧丽 [7] 对论文创新性评价的方法进行了综述，归纳出引文指标方法包括基于引用关系 [8]、基于组合引用 [9] 和基于引文动态网络的测度 [10]；内容指标方法包括基于关键词或主题词 [11]、基于引用内容 [12] 和基于机器学习的测度 [13-14]。汪雪峰等 [15] 则从论文新颖性和创新影响两个维度综合评价论文创新质量，弥补单一维度评价的不足。与非临床研究论文不同，临床研究论文的评价更强调研究设计的严谨性和结果的临床相关性。这一特点使得临床研究论文的评价标准与传统论文存在显著差异。目前针对非临床研究论文的评价研究已有大量文献且有最新综述 [7]，因此本文将重点对现有高质量临床研究论文识别的相关研究进行梳理。

本文首先采用文献调研法对临床研究及临床研究论文质量的相关概念与内涵进行梳理，然后系统搜集了国内外有关高质量临床研究论文自动识别方法的相关文献，以 PubMed 和中国知网为数据源，PubMed 收录论文检索式为（“deep learning”[Title/Abstract] OR “machine learning”[Title/Abstract] OR “automatic identification” [Title/Abstract] OR “auto-identify” [Title/Abstract] OR “information retrieval” [Title/Abstract])AND (“clinical research”[Title/Abstract] OR “clinical trial*”[Title/Abstract] OR “clinical article*”[Title/Abstract] OR “clinical evidence*” [Title/Abstract]) AND (“high quality”[Title/Abstract] OR “scientifically rigorous” OR “high level” [Title/Abstract])，中国知网收录论文检索式为 SU=(‘深度学习’+‘机

器学习’+‘自动识别’) and SU=(‘临床研究’+’临床试验’+‘临床论文’+‘临床证据’）and SU=(‘高质量’+‘科学严谨’+‘高水平’)。经过人工筛选，目前尚未有自动识别高质量临床研究论文的中文文献。为此，本文对检索获取的英文文献进行了引文追溯，补充了漏选文献，对相关英文文献进行重点分析。通过阅读相关文献，发现高质量论文主要集中在高质量临床研究论文金标准的建立和识别方法等方面。因此，本文将从这两个方面进行总结和归纳。

1 相关概念与内涵

1.1 临床研究的概念

当前关于临床研究的概念尚未达成共识，总体上分为广义的临床研究以及狭义的临床研究。国际权威机构或组织对于临床研究的界定是从广义角度出发：

（1）美国国立卫生研究院（National Institutes of Health，NIH）：将临床研究定义为目的在于了解疾病并改善人类未来医疗保健手段的涉人医学研究，具体类型包括[16]：

①流行病学研究，即通过研究特定人群健康和疾病的模式、病因和影响因素等来提高对疾病的认知。

②行为学研究，即研究人类行为及其与健康和疾病的关系，以提高对此方面的认识。

③卫生服务，着眼于人们获得医疗保健提供者和医疗保健服务的方式、所花费的医疗成本以及患者就诊结局。

④临床试验，评估干预措施对健康结局的影响。

（2）美国 clinicaltrials.gov：定义临床研究为针对人类志愿者(也称为参与者)开展的研究，旨在产生医学知识，主要有临床试验（也称为干预性研究）和观察性研究两种类型[17]。

（3）加拿大卫生研究院：将临床研究界定为以改善疾病和损伤的诊断和治疗（包括康复和姑息治疗）以及提高个体在正常生命历程中的健康和生活质量为目标的研究，包括以患者为对象或以治疗患者为目的而开展的研究[18]。

（4）法国国家健康与医学研究院 (Institut national de la santé et de la recherche médicale，Inserm)：定义临床研究为以生产生物学或医学知识为目的对人类受试者开展的研究。包括对新药、医疗器械或生物制剂进行评估的临床试验；识别与疾病有关的分子或细胞机制，用于发现新的治疗靶点；为实施预防策略，识别相关的遗传或环境风险因素；比较现有的诊断或治疗方法，用于制定诊疗指南等[19]。

（5）中华医学会：组织临床流行病学、临床医学、统计学、医学科研管理学等多学科专家对临床研究的基本特征进行了广泛深入的研讨，形成共识，认为临床研究是以疾病的诊断、治疗、预后和病因为主要研究内容，以患者为主要研究对象，以医疗服务机构为主要研究场所，由多学科人员共同参与组织实施的科学研究活动[20]。

基于上述内涵界定，可见广义的临床研究是医学健康研究的组成部分，旨在产生认知、预防和治疗人类疾病以及促进健康等方面有价值的知识，具体包括干预性临床研究和观察性临床研究。其中干预性临床研究又称为临床试验，即为狭义的临床研究。如美国 NIH 癌症术

doi:10.3772/j.issn.2095-915x.2025.03.004

语词典（NCI Dictionary of Cancer Terms）中对于临床研究的界定也是从狭义角度出发，定义临床研究为一种测试筛查、预防、诊断或治疗疾病的新方法在人体中效果的研究[21]。本研究是从广义的临床研究内涵出发，综述国内外自动识别高质量临床研究论文的研究进展。

1.2 临床研究论文质量的内涵

国内外学者主要是从报告质量和研究自身质量两个维度来定义临床研究论文的质量。1981 年第一个随机对照研究质量评价量表 Chalmers TC 量表问世[22]，随后陆续产生了数十种量表与清单，可用于临床研究论文质量评价，涉及报告质量或研究质量。

1994 年 Goodman 等[23]将临床研究论文质量从概念上明确分为研究本身的质量和研究报告的质量，其中将研究报告质量描述为作者对研究的陈述是否足够详细和清晰，以便读者可以判断其数据和结论的优缺点。Moher 等[24]于 1995 年鉴定了评价随机对照试验（Randomized Controlled Trial，RCT）质量的 9 种清单和 60 余种量表，将研究质量定义为在对照试验设计、实施和分析过程中，应尽可能避免偏倚产生或采取措施将偏倚控制在最低程度，侧重于方法学质量；将报告质量定义为对试验设计、实施和分析等信息的报告情况。该研究发现已有的质量评价工具对于研究质量的评价有局限性，可能会受报告质量的影响。基于此，Jadad[25]进一步将研究质量定义为试验设计产生无偏结果、接近真实情况的可能性，并开发了著名的“Jadad 量表”用于评价方法学质量，之后成为评价 RCT 质量的重要工具。随后，Verhagen 等[26]指出 Jadad 的定义中只涵盖了内部真实性，并通过开展三轮德尔菲调查进一步将研究质量定义为反映结论有效性的相关因素，包括内部真实性、外部真实性和统计分析等。但王家良[27]提出，内部真实性包含了统计学方法，外部真实性与临床重要性表述可互换，内部真实性、偏倚风险和方法学质量基本等价。总体上，评价研究自身质量的核心是其方法学质量，即对研究的真实性、风险偏倚、内部效度的评价。

循证医学中，系统综述、Meta 分析等二次研究以及指南制定均需要在研究设计阶段确定原始研究的纳入排除标准和质量评价方法，并报告对原始研究的质量评价结果，作为判定证据可信度的依据，这里的评价结果称为证据质量，也主要是从真实性和有效性等方法学方面（比如随机、盲法、偏倚等）界定原始研究的质量[28]。

针对文献的报告质量和方法学质量，多个组织已经制定了国际公认的、面向不同研究设计的评价工具，包括量表、清单或条目[29]。报告质量评价方面，形成了 RCT、诊断试验、观察性研究和非随机对照试验（Non-randomized controlled trial，NRCT）相应的报告规范 CONSORT[30]、STATD[31]、STROBE[32] 和 TREND[33]。方法学质量方面，发布了多个偏倚风险评估工具如 ROB2.0（RCT）[34]、ROBINS-I (NRCT)[35]、QUADS（诊断试验）[36]、NOS（观察性研究）[37]、JBI（现况调查及分析性横断面研究）[38] 等。循证医学中，还形成了对原始研究开展的证据质量评价工具，如 GRADE[39] 等。针对上述工具，量表可用于计算所有条目的总分，清单只能逐个条目进行比对

评价。现有研究主要是采用人工评价的方式选择不同的评估工具基于不同的目的对临床研究论文的质量进行评价[40]，但目前还是存在误用方法学质量和报告质量评价工具的问题，如使用 CONSORT 来评价 RCT 的方法学质量。

本文认为临床研究论文质量的界定与其评价目的有关，取决于是从审稿角度还是从循证角度进行评价。从审稿角度出发，除了论文方法学质量，还要注重报告质量；但从循证角度出发，更应关注方法学质量，即应突出研究是否接近真实或真值。

2 高质量临床研究论文的金标准

本文基于循证目的，从辅助提高临床医生获得有效证据的效率出发，分析了自动识别高质量临床研究论文的相关研究。通过系统分析现有文献发现，建立自动识别方法的关键之一是有金标准数据集。以往研究采用的金标准包括来自 Clinical Hedges 和《美国内科医师学会杂志俱乐部》（ACP Journal Club）的论文集、临床指南引用论文，以及 Cochrane 系统综述引用论文。

2.1 Clinical Hedges数据库

Clinical Hedges 数据库[41]由麦克马斯特大学开发，包含 170 种临床期刊于 2000 年发表的论文，其中 161 种期刊被 MEDLINE 收录，135 种被 EMBASE 收录。由专家和高级校准人员识别并标记每篇论文为与人体健康照护相关的原始研究或综述性研究以及归于何种研究目的，具体包括人类疾病的病因、预后、诊断、预防或治疗或康复、临床预测或卫生经济学，以及卫生服务质量改进、卫生人员的继续教育和定性研究等，如表 1 所示；除了定性和成本研究外，其他目的的研究均基于预先制定的临床研究方法学标准，被人工标记为“达标”或“不达标”，“达标”的研究即视为高质量临床研究。该数据库是基于对不同目的临床研究所采用研究方法的严谨性来评价临床研究的质量。通过该方法形成的数据集已被 MEDLINE、EMBASE 等数据库作为金标准用于制定最佳检索策略[42]。

2.2 《美国内科医师学会杂志俱乐部》

《美国内科医师学会杂志俱乐部》总结了来自 120 多种内科临床期刊的最佳新证据[45]。麦克马斯特大学的研究人员和临床领域编辑人员严格评估医学文献的科学价值，由 3600 多名医生组成的全球专家小组评估具有学术严谨性的研究的临床相关性和新闻价值。目的是使用预先规定的方法学标准从生物医学文献中标记出有关内科疾病的病因、病程、诊断、临床预测、预防、治疗或经济学以及质量改进和继续医学教育干预试验的最佳原创性研究和系统综述，并在“结构化摘要”中总结这些原始研究和综述。由对论文主题领域专业水平较高且声明没有利益冲突的临床医生对每篇论文的研究结果的背景、方法和临床应用进行简要评论。这类论文值得医生及时关注，以紧跟内科学重要进展。

原创或综述性论文的基本标准为：用英文撰写；涉及人类；涉及对医学、护理、康复和其他卫生专业的临床实践很重要的主题，而不是对疾病流行率的描述性研究；分析方法与研究问题相一致。此外，针对不同研究目的的论文，附加了方法学标准，如表 2 所示，依据此标准

doi:10.3772/j.issn.2095-915x.2025.03.004

表 1 临床原创或综述性论文的研究目的分类及高质量标准 [43-44]

目的	定义	方法严谨性标准
病因	与确定暴露与疾病或症状之间是否直接相关的研究内容。例如，“导致人类罹患疾病的原因？”	关注暴露和临床结局之间的关系；数据收集是前瞻性的；有明确的比较组；对暴露结局的观察者设盲。
预后	与研究开始时即存在的疾病或症状的临床进程或自然过程的预测直接相关的研究内容。	队列之初不存在感兴趣的临床结局；截至出现主要研究终点或研究结束时随访 80% 的患者；研究分析与研究设计相一致。
诊断	涉及使用工具来诊断疾病或症状的直接相关的研究内容。	纳入一系列参与者；有客观诊断标准（金标准）或当前临床诊断标准；参与者接受新型检测和某种形式的诊断标准；在未知检测结果的情况下解释诊断标准，反之亦然；分析符合研究设计。
治疗	直接与治疗干预（包括不良反应研究）、预防、康复、质量改进或继续医学教育有关的研究内容。	将参与者随机分配到比较组；在任何既定的随访评估中，至少 80% 参与者的结局评估纳入 1 项主要分析；分析与研究设计一致。
成本	直接与医疗保健问题的成本、融资或经济相关的研究内容。	—
经济学	与医疗保健问题的经济学直接相关（成本论文的子集）的研究内容。	研究问题是比较替代方案；对替代服务或活动在产生的效果（有效性）和消耗的资源（成本）方面进行比较；有效性的证据必须来自符合上述诊断、治疗、质量改进标准的真实患者研究或系统综述文章；基于个体患者数据的有效性和成本估算（微观经济学）；结果以一种干预措施相对于另一种干预措施的增量或额外成本和有效性形式呈现；如果存在不确定性，则进行敏感性分析。
临床预测指南	直接与疾病或症状某些方面的预测有关的研究内容。	指南是在一组或多组真实患者（训练集）中形成的；指南在另一组真实患者（测试集）中得到验证。
定性	与人类在某些情况下的感受或体验有关的研究内容，数据收集方法和分析适用于定性数据。	—
其他	研究内容不符合上述任何定义。	—

表 2 《美国内科医师学会杂志俱乐部》中关于不同研究目的高质量文献的筛选附加标准 [45]

目的	附加标准
预防或治疗	● 将参与者随机分配到比较组 ● 每组≥ 10 名患者完成主要结局评估 ● ≥ 80% 的随机受试者在规定的随访点进行了≥ 1 项特定结局的评估 ● 测量已知或可能具有临床重要性的结局 ● 亚组分析必须预先计划，按随机分组进行分组分析；必须分析≥ 2 个亚组之间的相互作用。
诊断	● 纳入一系列参与者，所有参与者均疑似患有该病，其中一些患者（但不是全部）在诊断检测后发现患有目标疾病 ● 纳入≥ 100 名参与者，其中≥ 50 名参与者患病，≥ 50 名参与者未患病 ● 有客观诊断标准（金标准）（例如，实验室检测）或当前的临床诊断标准（例如，深静脉血栓的静脉造影），最好有可重复标准的文件用于主观解释诊断标准（即报告观察者之间超越偶然性之外的存在一致性的统计学显著性衡量标准） ● 每个参与者都必须接受新型检测和某种形式的诊断标准 ● 在不知道检测结果的情况下解释诊断标准 ● 在不知道诊断标准结果的情况下解释检测 诊断检测也可以在随机试验中进行测试，在这种情况下，预防或治疗标准适用于此。 鉴别诊断研究必须满足以下附加标准： ● 一群出现相似的、最初未被诊断但可重复定义的临床问题的患者 ● 临床环境的明确描述，包括转诊过滤（the referral filter） ● 对所有患者使用可重复的诊断检查策略，确定≥ 80% 患者的诊断，并进行随访直至患者确诊或随访≥ 1 个月（急性疾病）或≥ 1 年（慢性或复发性疾病）。

（续表）

目的	附加标准
预后	● 初始队列的患者具有相似的疾病且处于早期阶段，所有患者最初都没有感兴趣的临床结局 ● 基于研究目的（即识别和跟进初始队列）的前瞻性标准化数据收集 ● 随访≥ 80% 的患者，直至主要研究终点发生或研究结束。
临床预测指南	● 目的是验证或比较规则 / 指数 / 量表 / 模型，该规则 / 指数 / 量表 / 模型将≥ 2 个因素组合成某种类型的评分 / 排名，该评分 / 排名基于存在 / 不存在这些因素，将患者分配为特定结局（诊断、预后、治疗）的不同风险水平 ● 预测指南的数据必须在其预测的结局之前即可用 ● 该指南必须在一组或多组真实（非假设）患者（衍生或开发队列）中产生 ● 该指南必须在另一组真实（非假设）患者（验证队列）中进行验证；不允许内部推导作为验证 验证先前推导的临床预测指南的研究必须为该指南的推导提供参考，并且必须评估衍生队列中≥ 1 项的结局 使用≥ 1 项研究的个体患者数据开发的预测指南不需要单独验证 ● 研究必须提供有关如何在患者个体中应用预测指南或对此信息提供参考内容。
医疗干预危害的病因学	● 明确研究目的是评估干预措施的有害影响 ● 前瞻性收集数据，明确确定比较组（按偏好降序排列：半随机对照试验、非随机对照试验、通过逐案匹配或统计调整来创建可比较组的队列研究、巢氏病例对照研究） ● 对暴露结局的观察者设盲（掩蔽）（如果结局是客观的，例如全因死亡率或客观检测，则假设满足标准） ● 如果报告有危害，则相对风险（RR）或风险比（HR）或等效值≥ 2.0 且其 95% CI 的下限高于 1.5 ● 如果没有报告有危害，则 RR 或 HR 或等效值的 95% 置信区间上限低于 1.5 评估不良反应的随机对照试验使用预防或治疗研究的标准进行评估。
质量改进或继续教育	● 将参与者或参与对象随机分配到对照组 ● 每组≥ 10 名患者（评估结局） ● ≥ 80% 的随机受试者在规定的随访点进行了≥ 1 项的主要结局评估 ● 测量已知或可能的对临床或教育重要的结局。
医疗保健计划或干预措施的经济学研究	● 解决的经济问题必须基于对真实患者的替代方案的比较 ● 替代诊断或治疗服务或质量改进活动必须根据产生的效果（有效性）和消耗的资源（成本）进行比较 ● 在通过预防或治疗标准的单一随机对照试验中报告了有关有效性和成本的证据 ● 结果应以一种干预相对于另一种干预的增量或额外成本和结局来呈现 ● 如果预测存在不确定性或测量不精确，则应进行敏感性分析。
系统综述	● 明确陈述临床主题 ● 方法学描述，包括检索的数据库以及论文的纳入和排除标准；关于预后的综述必须有“初始队列”作为纳入标准 ● 检索了≥ 1 个主要数据库 明确不包括≥ 1 项符合上述治疗、诊断、鉴别诊断、预后、临床预测指南、医疗干预危害的病因学、质量改进或经济学标准的综述被排除。
个体患者数据荟萃分析	● 包括来自≥ 2 项关于预防、治疗、诊断、鉴别诊断、预后、临床预测、病因学、质量改进或医疗保健经济学的原始研究的患者水平的组合数据。

进一步筛选最高质量的研究文献。若后续发现可应用更高的标准来提高临床实践研究的有效性和适用性，则相关标准会被修改。与 Clinical Hedges 数据库类似，美国内科医师学会杂志俱乐部也是对不同研究目的的论文，制定了方法学标准，以此为指标评价临床研究论文的质量。

《美国内科医师学会杂志俱乐部》只选择了内科学领域相关期刊，在领域应用上具有一定局限性。相比于 Clinical Hedges 数据库，该数据集对研究目的的分类略有差异，筛选高质

doi:10.3772/j.issn.2095-915x.2025.03.004

量文献的标准更加细致和严格，在预防或治疗、诊断、质量改进或继续教育等研究目的下明确了患者的最低样本量。

2.3 临床指南引用论文

临床指南中包含有关特定疾病诊断和治疗的循证建议。基于严格的系统评价方法，领域专家确定所有与临床指南主题相关的研究，筛选出符合标准的研究，对研究进行证据质量分级，将证据质量、患者偏好与价值观以及卫生经济成本等进行综合考虑，形成指南推荐强度等级意见[39, 46]。Bian 等[47]研究提取了心力衰竭等疾病的 11 个最新版本的指南中治疗部分引用的论文（包括 RCT、系统综述和 Meta 分析）作为金标准，其中要求指南必须满足：指南中引用的论文必须通过系统的搜索、筛选和评估进行筛选；符合条件的指南必须提供明确的治疗建议，以及对支持每项建议的原始研究的引用。虽然临床指南引用的论文在指南制定过程中，由专家进行了一定的筛选，但最终临床指南中纳入的证据质量依然有高低之分。所以将临床指南引用的所有论文作为衡量高质量临床研究论文的金标准存在一定的局限性。该类金标准的确立依据是认为经过专家系统评估后符合临床指南纳入证据的临床研究论文即为高质量临床研究。

2.4 Cochrane系统综述数据库

Cochrane 协作组织是一个国际组织，主要任务是通过系统汇总、评价和分析临床研究的证据，为临床实践提供可靠、科学的指导。该组织主要产品为 Cochrane 图书馆（Cochrane Library），包含了为医疗保健决策提供的各种证据，Cochrane 系统评价数据库（Cochrane Database of Systematic Reviews，CDSR）是其高质量证据库之一[48]。Cochrane 系统综述使用透明和系统的过程，针对某一具体的研究问题，全面收集相关研究，用统一、科学的标准筛选合格研究，评估其质量，并综合定性或定量的结果，得出可靠结论[49]。Cochrane 系统综述在全球范围内被公认为循证医学最高级别的证据，是临床决策的最佳证据来源，由于其在开展过程中会对纳入的文献进行研究质量的评价，确定证据等级和推荐强度，已有研究[2]利用 Cochrane 系统综述中引用的研究作为高质量研究的代表。但由于系统综述涉及依赖专家筛选文献，且综述发表也需要一定周期，因此 Cochrane 系统综述通常不能纳入最新发表的研究[2]。与临床指南引用论文类似，Cochrane 综述引用的论文也被认为是通过专家评估后的高质量论文。

综上，以上 4 个金标准数据集是从不同的角度来定义高质量临床研究。Clinical Hedges 和《美国内科医师学会杂志俱乐部》对高质量临床研究的界定相似，是从研究本身的目的出发，从研究设计到具体开展，分析整个过程的科学性、严谨性，达到一定标准的研究才可视为高质量的临床研究，即对应 1.2 部分内涵界定中的方法学质量高。指南引用论文和 Cochrane 系统综述引用论文是从指南和综述开展的严格性和科学性角度出发，认为指南和综述引用的论文都是人工筛选过的具有一定质量要求的研究论文，将其等价为高质量临床研究，但实际上所纳入的研究质量还是有高低之分。

3 高质量临床研究论文自动识别方法

分析现有关于高质量临床研究论文自动识别方法，从发展路径来看，最初是通过半自动化的检索过滤器来进行识别，随后逐步引入了基于传统机器学习的识别方法，近年来则发展为基于深度学习技术进行识别。

3.1 检索过滤器

检索过滤器（Search Filters）是通过制定检索策略，将其集成到检索工具中，实现高质量临床研究检索的功能。大多数检索过滤器是依赖于已标记了临床类别和方法学严谨性的文献数据库开发，并通过检测阳性（目标）和阴性（脱靶）论文的识别性能进行测试[50]。如Haynes 等[42, 50]开发最佳的 MEDLINE 搜索策略，以检索出为预防或治疗疾病等不同目的开展的科学合理的临床研究。基于 Haynes 的一系列研究，已经开发了 5 个主题（即治疗、诊断、病因学、预后、临床预测指南）的 PubMed 临床查询过滤器（PubMed Clinical Query Filter），具体如表 3 所示。该过滤器已被用作验证从 PubMed 检索高质量临床研究的新方法有效性的基础。

这类方法的关键之处在于构建一套全面的检索词，即列出满足不同研究目的标准的相关检索词。例如，对于 MEDLINE，利用医学主题词（MeSH），包括出版物类型（pt）和副标题（sh），或者文本词（tw），然后通过访谈、会议、电子邮件方式以及向美国国家医学图书馆提出咨询，用于征求临床医生和图书馆员的意见。以 Clinical Hedges 作为金标准，将通过检索策略检索到的研究与 Clinical Hedges 中人工标识的高质量临床研究进行比较以验证所制定的检索策略的性能。首先是确定每个术语和术语组合的敏感性、特异性、精确性和准确性。只有针对每种目的的研究检索出的灵敏度 >25% 和特异性 >75% 结果的单个搜索词才被保留，用于组合形成涵盖两个及以上检索词的搜索策略。为了开发多项检索策略以优化灵敏度或特异性，研究组测试了所有灵敏度≥ 75% 且特异性≥ 50% 的双项检索策略；为了优化准确性，考虑了准确性 >75% 的双项检索策略。最终形成了面向不同研究目的的 PubMed 临床查询过滤器，提供高特异性和高灵敏度两种检索方式，使用者可以根据具体搜索需求权衡灵敏度和特异性。高特异性限制了检索到的非目标论文的比例，有助于临床医生高效地快速查找到相关证据；高灵敏度检索用于识别所有可能相关的论文，但会包括更高比例的非目标论文[51]，这种方式更适合用于系统综述的文献检索。

PubMed 临床查询过滤器是通过使用人工标注的面向不同研究目的的高质量临床研究作为金标准所开发的检索策略来检索出高质量的临床研究论文。尽管检索过滤器有效提高了高质量临床研究文献的检索效率，但仍存在局限性。首先，该方法依赖于 MeSH 术语，然而 PubMed 平台有部分论文未被 MEDLINE 索引，关于临床试验的论文也需要时间进行人工索引确保质量[52]，因此未标注 Mesh 术语的论文会被遗漏。其次，PubMed 临床查询过滤器对诊断类研究识别不全，高特异性的检索过滤器会导致部分证据的遗漏[53]，而高灵敏度的检索过

doi:10.3772/j.issn.2095-915x.2025.03.004

滤器虽能减少筛选工作量，但仍需进行人工筛选[54]。总体而言，检索过滤器为临床研究文献检索提供了重要支持，但其效果依然需要根据具体应用场景进行权衡。

表 3 PubMed 临床查询过滤器中不同类型研究的具体检索策略[52]

分类	优化指标	灵敏度 / 特异度	PubMed 检索策略
治疗	高灵敏度	99%/70%	((clinical[Title/Abstract] AND trial[Title/Abstract]) OR clinical trials as topic[MeSH Terms] OR clinical trial[Publication Type] OR random*[Title/Abstract] OR random allocation[MeSH Terms] OR therapeutic use[MeSH Subheading])
	高特异度	93%/97%	(randomized controlled trial[Publication Type] OR (randomized[Title/Abstract] AND controlled[Title/Abstract] AND trial[Title/Abstract]))
诊断	高灵敏度	98%/74%	(sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnose[Title/Abstract] OR diagnosed[Title/Abstract] OR diagnoses[Title/Abstract] OR diagnosing[Title/Abstract] OR diagnosis[Title/Abstract] OR diagnostic[Title/Abstract] OR diagnosis[MeSH:noexp] OR (diagnostic equipment[MeSH:noexp] OR diagnostic errors[MeSH:noexp] OR diagnostic imaging[MeSH:noexp] OR diagnostic services[MeSH:noexp]) OR diagnosis, differential[MeSH:noexp] OR diagnosis[Subheading:noexp])
	高特异度	64%/98%	(specificity[Title/Abstract])
病因	高灵敏度	93%/63%	(risk*[Title/Abstract] OR risk*[MeSH:noexp] OR (risk adjustment[MeSH:noexp] OR risk assessment[MeSH:noexp] OR risk factors[MeSH:noexp] OR risk management[MeSH:noexp] OR risk taking[MeSH:noexp]) OR cohort studies[MeSH Terms] OR group[Text Word] OR groups[Text Word] OR grouped [Text Word])
	高特异度	51%/95%	((relative[Title/Abstract] AND risk*[Title/Abstract]) OR (relative risk[Text Word]) OR risks[Text Word] OR cohort studies[MeSH:noexp] OR (cohort[Title/Abstract] AND study[Title/Abstract]) OR (cohort[Title/Abstract] AND studies[Title/Abstract]))
预后	高灵敏度	90%/80%	(incidence[MeSH:noexp] OR mortality[MeSH Terms] OR follow up studies[MeSH:noexp] OR prognos*[Text Word] OR predict*[Text Word] OR course*[Text Word])
	高特异度	52%/94%	(prognos*[Title/Abstract] OR (first[Title/Abstract] AND episode[Title/Abstract]) OR cohort[Title/Abstract])
临床预测指南	高灵敏度	96%/79%	(predict*[Title/Abstract] OR predictive value of tests[MeSH Terms] OR score[Title/Abstract] OR scores[Title/Abstract] OR scoring system[Title/Abstract] OR scoring systems[Title/Abstract] OR observ*[Title/Abstract] OR observer variation[MeSH Terms])
	高特异度	54%/99%	(validation[Title/Abstract] OR validate[Title/Abstract])

3.2 机器学习方法

随着生物医学文献和电子健康记录的快速增长，基于机器学习方法的文本挖掘技术已被应用于文本摘要[55]、文献检索[56]和证据分级[57]，也被应用于系统综述的自动化筛选过程，识别相关论文，以减少科研人员工作量，提高科研效率[54]。此外，机器学习方法在生物医学文献中的另一个重要应用是改善临床相关论文的检索，一定程度上能克服布尔检索的局限性。下面将从传统机器学习和深度学习两个方面来分析其在识别高质量临床研究中的应用。

3.2.1 传统机器学习方法

多项研究通过构建基于不同研究目的的识别模型，来自动化筛选和识别高质量的临床研究论文。如 Marshall 等[58]应用传统机器学习方法来识别 RCT 研究。此类识别方法的过程通常包括几个关键步骤：确定金标准数据集、数据预处理、特征选择、模型构建以及性能评估。

（1）金标准数据集。金标准数据集是构建和评估机器学习模型的基础，本文第二部分总结的4种数据集是以往研究中使用过的金标准，包含专家标注过的原始研究和系统综述。即基于专家标注时采用的质量和临床相关性标准，遴选出是否符合评价标准的论文集用于训练机器学习模型。如 Aphinyanaphongs 等 [59] 利用《美国内科医师学会杂志俱乐部》论文作为金标准，训练机器学习模型，用于识别内科学高质量临床研究。Bian 等 [60] 应用临床指南治疗部分引用的论文作为金标准，构建模型自动识别高影响力的临床研究。

（2）数据预处理。数据预处理是机器学习模型的基础步骤之一，通常包括将文本转换为小写、提取词干和删除停用词等。Aphinyanaphongs 等 [59] 应用了术语加权方法，根据单词的出现频率来衡量每个词在文献中的重要性。Afzal 等 [57] 通过删除特定主题的常用术语和极少出现的术语来进行词表修剪。

（3）特征选择。特征选择决定了模型所依赖的信息种类，包括论文标题和摘要以及其他外部特征等。Aphinyanaphongs 等 [59] 使用 MeSH 术语、出版物类型以及论文标题和摘要等特征。Kilicoglu 等 [3] 采用了 MEDLINE 元数据、SemRep 提取的语义三元组、从标题和摘要中提取的 UMLS 概念和语义关系。Bian 等 [47] 构建的机器学习模型使用了文献计量指标、社交媒体关注和 MEDLINE 元数据等特征，具体包括期刊影响因子、样本量、基金支持、作者数量、临床有意义的句子数量、作者机构的学术影响、参考文献数量、论文页数、是否在 ClinicalTrials.gov 注册、是否发表在 PubMed Central 等。Lin 等 [61] 选择了期刊影响因子、研究设计、人类参与情况、样本量、P 值、置信区间。

（4）模型构建及评估。以往研究构建了不同的传统机器学习模型，包括支持向量机（Support Vector Machine，SVM）、朴素贝叶斯（Naive Bayes，NB）、决策树（Decision Tree，DT）、k 最近邻、随机森林等，其中 SVM 是应用最广泛的分类器；模型的验证方法通常使用 k 折交叉验证（如 10 折或 5 折），以确保模型的泛化能力；常用的性能评估指标包括敏感性（召回率）、特异性、准确性、曲线下面积（AUC）、F 值和精确度，Lin 等 [61] 还使用 Cohen’s kappa 系数来衡量模型预测结果与金标准之间的一致性。

通过对比以往采用传统机器学习作为识别方法的相关研究发现，将不同的特征进行组合 [2, 56]，可以提高传统机器学习模型的性能，例如，结合 MeSH 术语、文献计量特征等多种信息能够增强模型的识别能力。但是 MeSH 术语等特征依赖于 MELINE 的索引，PubMed 中未被索引到 MeSH 术语的论文不能被识别。此外，联合使用多个分类器可以平衡召回率和精确度 [3, 59]，但缺乏可解释性。

综上所述，传统机器学习方法通过学习专家标注的高质量临床研究论文特征，构建了用于自动识别新论文的模型。尽管基于传统机器学习的识别方法很有效，但也具有较大的局限性，如对重要特征工程（例如，UMLS 概念和语义关系、文献计量指标）的要求、对时间相关特征（例如，MeSH 术语、引用频次）的依赖，以及对特定数据集的潜在过度拟合等，导致无法在实践中广泛应用。

doi:10.3772/j.issn.2095-915x.2025.03.004

3.2.2 深度学习方法

深度学习是机器学习的一个重要分支，通过多层神经网络来学习复杂数据的多级抽象表示。不同于传统机器学习依赖人工设计的特征，深度学习能够自动从大规模数据中提取特征。深度学习的核心机制是反向传播算法，通过逐层调整模型参数，发现数据中的复杂结构，从而提升分类性能。Del Fiol 等 [56] 以标题和摘要作为特征，使用 PubMed 临床查询治疗过滤器检索的数据集训练和测试卷积神经网络（Convolutional Neural Network，CNN），以 Clinical Hedge 注释数据集作为金标准，与 PubMed 临床查询过滤器等先进的搜索过滤器相比，深度学习模型性能表现良好。Marshall 等 [58] 在 Cochrane Crowd RCT 数据集的标题和摘要上训练和优化了 SVM 和 CNN 来识别 RCT，使用 Clinical Hedges 数据集评估了模型性能，ROC 指标显示 CNN 模型优于 SVM。Afzal 等 [2] 基于 Cochrane 系统综述数据库中高质量研究以及 PubMed 查询的一般性研究，构建了多层感知器（Multilayer Perceptron，MLP）模型，并与浅层机器学习方法 NB、DT、SVM 和梯度提升树（Gradient Boosting Trees，GBT）进行模型性能比较，发现所建立的 MLP 模型表现最好。

与传统机器学习方法相比，深度学习具有多方面的优势。首先，它降低了对复杂特征工程的依赖，仅需标题和摘要等基本内容特征即可实现较好的分类性能。其次，深度学习方法在不依赖时间敏感特征（如 MeSH 术语和文献引用频次）的情况下，仍能取得较高的召回率和准确性。尽管使用文献的全文可能进一步提升模型性能，但由于部分期刊不开放全文访问，这在实践中存在一定困难。此外，相比于 PubMed 临床查询过滤器，深度学习方法表现良好，可以有更高的召回率及精度 [56,58]。

自然语言处理（Natural Language Processing，NLP）的最新进展，尤其是预训练语言模型（Pre-trained Languages Model，PTLM）的发展，进一步推动了深度学习在文本分类中的应用。BERT、GPT 等模型在处理文本上下文理解方面表现优异，BERT 尤其因其双向机制而被广泛应用于文本摘要、检索、问答等任务。Abdelkader 等 [62]、Lokker 等 [63] 构建 BERT、BioBERT、BlueBERT 和 PubMedBERT 模型，利用基于临床研究方法学严谨性手动标记了 2012—2020 年 160000 余篇 PubMed 论文，基于此进行学习和验证，最终确定 BioBERT 在识别高质量临床研究中具有最佳性能。

与传统机器学习相比，深度学习方法同样通过学习专家标注过的高质量临床研究论文来构建模型，但其主要依赖标题和摘要等内容特征，极大地减少了对文献计量指标等外部特征的依赖。然而，深度学习在处理类别不平衡问题时面临一定挑战，例如 RCT 研究数量显著少于非 RCT 研究，高质量临床研究数量也少于非高质量临床研究。为了解决这一问题，Marshall 等 [58] 和 Abdelkader 等 [62] 采用了平衡采样技术，充分利用负样本进行训练，生成多个模型，通过集成策略来提升最终分类性能。

综上，深度学习方法在自动识别高质量临床研究方面表现出显著优势，尤其是在复杂特征自动提取和模型性能上。然而，类别不平衡问题仍是其面临的主要挑战，如何有效处理此类问题将是未来研究的一个重要方向。

4 问题及启示

随着生物医学文献的快速增长，临床医生面临着跟踪最新知识和获取高质量证据的重大挑战。当前，依赖于人工评估的传统方法已经难以满足需求，亟需通过人工智能技术进行创新和改进。

4.1 亟需开发覆盖外科领域、涵盖最新临床研究的金标准数据

自动识别高质量临床研究的性能在很大程度上取决于所使用的金标准数据集。然而，当前关于高质量的临床研究金标准可选择性有限，且均存在一定局限性，特别是对外科领域和最新临床研究的覆盖不足，如 Clinical Hedges 数据库人工标注的文献发表于 2000 年，无法涵盖最新的研究成果；《美国内科医师学会杂志俱乐部》尽管建立了较为详细的标准，但其高质量论文主要局限于内科领域；指南引用论文和 Cochrane 综述引用论文虽然经过人工筛选，但仍存在低质量研究的可能。

对比现有金标准数据集对于高质量研究的定义，本文认为《美国内科医师学会杂志俱乐部》所建立的标准更为详细，更能筛选出高质量研究。因此，改进现有金标准数据集的路径包括：首先，以《美国内科医师学会杂志俱乐部》所采用标准为基础，进一步扩展到外科领域 [59]，明确外科领域的评价标准，形成覆盖外科领域的金标准数据集；其次，以 Clinical Hedges 数据库为基础，依据注释标准补充并更新最新的临床研究，确保金标准数据能够反映最新的医学进展。

未来，随着人工智能技术的进步，金标准数据集的注释工作可以逐步实现自动化，减轻人力负担，确保数据集的更新速度和覆盖广度。

4.2 亟待拓展面向非英语语种高质量临床研究的自动识别模型

当前，高质量临床研究的自动识别主要集中在英语语种发表的论文，针对非英语语种的研究识别较为稀缺。国内学者大多从研究设计科学性 [64]、报告质量 [65-66]、伦理规范 [67] 等角度对中文临床研究论文质量进行人工评价。栾嘉等 [40] 从医学期刊审稿角度出发，构建了包括伦理、真实性、实用性、透明性、科学性、创新性 6 个维度的临床研究论文质量评价体系，但该体系依然依赖人工评价。总之，国内缺少自动识别模型，当然由于缺乏中文高质量临床研究数据集作为自动识别模型学习的语料，也限制了自动识别模型相关研究的开展。

科学有效的临床实践指南离不开本土临床研究证据，照搬套用他国指南可能会对本国患者带来一定危害。2016 年 5 月 30 日，习近平总书记在“科技三会”上提出广大科技工作者要把论文写在祖国的大地上。2020 年 2 月，科技部印发的《关于破除科技评价中“唯论文”不良导向的若干措施（试行）》中指出“发表在具有国际影响力的国内科技期刊”作为“三类高质量论文”中的一类。中国的科技评价导向将指引研究者把高水平论文发表在国内高水平期刊。因此识别中文高质量临床研究论文对于制定本土化的临床指南具有重要意义。

对我国来说，一方面可以构建专门针对中

doi:10.3772/j.issn.2095-915x.2025.03.004

文的高质量临床研究数据集，并结合机器学习和 NLP 技术，开发本土的自动识别模型；另一方面，开发跨语言的 NLP 技术，如多语言版本的 BERT、GPT 等大模型，进一步提升非英语语种论文的自动识别能力。这将助力国内相关文献数据库实现识别高质量临床研究的功能，从而为本地临床医生提供更加适用的自动化工具，助力其获取高质量证据。

4.3 亟需完善识别高质量临床研究的人工智能方法

目前用于识别高质量临床研究的自动化方法存在一定局限性。PubMed 临床查询过滤器在识别诊断类研究时效果不佳，高特异性的检索过滤器虽然可以减少不相关文献的干扰，但会导致部分证据缺失，而高灵敏度的检索过滤器虽然能够增加召回率，却无法完全替代人工筛选文献的工作[53-54]。传统机器学习方法依赖复杂的特征工程，且容易受时间相关特征（如 MeSH 术语、引用频次）的影响，并可能出现过拟合问题。

近年来，NLP 领域的快速发展，特别是 GPT 和 BERT 等预训练语言模型的兴起，为临床研究识别提供了新的可能。BERT 已在多项任务中展现了其卓越的性能，但在识别高质量临床研究中，尚缺乏与 CNN 等深度学习模型以及 GPT 等大语言模型的对比研究。结合强化学习与元学习等方法可以进一步提高模型在新任务或数据集上的适应性。强化学习通过奖励机制优化识别决策，而元学习可以帮助模型快速适应不同研究领域或类别的任务。此外在类别不平衡和某些类别（例如预后或诊断类）论文缺乏的处理方面还有待优化，未来，可利用数据增强技术，如生成式对抗网络（Generative Adversarial Network，GAN）或数据合成方法，以增加少数类别的数据样本，进而提升模型在不平衡数据集上的表现。

高质量临床研究识别的未来发展方向将集中在跨语言、跨类别、多模型融合与多任务学习方面。通过结合多种深度学习模型及先进的数据增强技术，进一步提升识别的全面性与准确性。将性能优异的自动识别模型集成到现有的文献数据库中，以有效提高临床医生获取高质量证据的效率，并推动医学领域的循证实践。

5 结语

随着生物医学文献的迅速增长，临床医生面临着获取最新知识和高质量证据的重大挑战。传统的依赖人工评估的方法已无法满足这一需求，因此亟需引入人工智能技术进行创新与改进。本文首先明确了临床研究及其论文质量的相关概念，然后围绕高质量临床研究论文的识别技术体系，系统回顾了国际上从金标准到具体识别技术的相关研究成果。为此，建议未来进一步开发覆盖外科领域和最新临床研究的金标准数据集，扩展非英语语种高质量临床研究的自动识别模型，并优化识别高质量临床研究的人工智能方法。

参 考 文 献

[1] MEDLINE PUBMED PRODUCTION STATISTICS. National Library of Medicine[EB/OL]. (2024-4-30) [2025-03-01]. https://www.nlm.nih.gov/bsd/medline_

pubmed_production_stats.html.

[2] AFZAL M, PARK B J, HUSSAIN M, et al. Deep learning based biomedical literature classification using criteria of scientific rigor[J]. Electronics, 2020, 9(8): 1253.

[3] KILICOGLU H, DEMNER-FUSHMAN D, RINDFLESCH T C, et al. Towards automatic recognition of scientifically rigorous clinical research evidence[J]. Journal of the American Medical Informatics Association, 2009, 16(1): 25-31.

[4] 王超越 . 基于引文网络和替代计量学的单篇学术论文影响力评价 [D]. 济南：山东师范大学 , 2024.

[5] 付伟棠 . 我国学术期刊同行评议研究综述 [J]. 中国科技期刊研究 , 2019, 30(8): 819-826.

[6] 罗卓然 , 王玉琦 , 钱佳佳 , 等 . 学术论文创新性评价研究综述 [J]. 情报学报 , 2021, 40(7): 780-790.

[7] 曹慧丽 . 学术论文创新性评价方法研究综述 [J]. 出版与印刷 , 2024(2): 71-78.

[8] KOSMULSKI M. Successful papers：a new idea in evaluation of scientific output[J]. Journal of Informetrics, 2011, 5(3): 481-485.

[9] UZZI B, MUKHERJEE S, STRINGER M, et al. Atypical combinations and scientific impact[J]. Science, 2013, 342(6157): 468-472.

[10] BORNMANN L, TEKLES A. Disruptive papers published in scientometrics[J]. Scientometrics, 2019(120): 331-336.

[11] 任海英 , 王德营 , 王菲菲 . 主题词组合新颖性与论文学术影响力的关系研究 [J]. 图书情报工作 , 2017, 61(9): 87-93.

[12] 姜霖 , 张麒麟 . 基于引文细粒度情感量化的学术评价研究 [J]. 数据分析与知识发现 , 2020, 4(6): 129-138.

[13] 李贺 , 杜杏叶 . 基于知识元的学术论文内容创新性智能化评价研究 [J]. 图书情报工作 , 2020, 64(1): 93-104.

[14] 曹树金 , 赵浜 , 岳文玉 , 等 . 学术论文创新点的识别与检索入口研究 —— 以情报学期刊论文为例 [J]. 现代情报 , 2021, 41(12): 17-27.

[15] 汪雪锋 , 于慧妍 , 郑思佳 , 等 . 学术论文创新质量评价研究 —— 以多能干细胞技术为例 [J]. 数据分析与知识发现 , 2024, 8(5): 127-138.

[16] NIH. NIH clinical research trials and you[EB/OL]. (2025-03-20) [2025-05-28]. https://www.nih.gov/health-information/nih-clinical-research-trials-you/basics.

[17] NCBI. Learn about studies[EB/OL]. (2024-06-10) [2025-05-28]. https://clinicaltrials.gov/ct2/about-studies/learn.

[18] CANADIAN INSTITUTES OF HEALTH RESEARCH. Clinical research[EB/OL]. (2019-05-15) [2024-03-01]. https://cihr-irsc.gc.ca/e/48805.html.

[19] INSERM. Clinical research[EB/OL]. (2021-02-07) [2024-03-01]. https://www.inserm.fr/en/our-research/clinical/clinical-research/.

[20] 赵一鸣 , 曾琳 , 李楠 . 临床研究的四个基本特征 [J]. 中华医学杂志 , 2012, 92(36): 2521-2523.

[21] NATIONAL CANCER INSTITUTE. Clinical research[EB/OL]. (2024-03-01) [2024-03-01]. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/clinical-research.

[22] CHALMERS T C, SMITH H JR, BLACKBURN B, et al. A method for assessing the quality of a randomized control trial[J]. Controlled Clincal Trials, 1981, 2(1): 31-49.

[23] GOODMAN S N, BERLIN J, FLETCHER S W, et al. Manuscript quality before and after peer review and editing at Annals of Internal Medicine[J]. Annals of Internal Medicine, 1994, 121(1): 11-21.

[24] MOHER D, JADAD AR, NICHOL G, et al. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists[J]. Controlled Clincal Trials, 1995, 16(1): 62-73.

[25] JADAD A. Assessing the quality of reports of randomized clinical trials: is blinding necessary?[J]. Controlled Clinical Trials, 1996, 17(1): 1-12.

[26] VERHAGEN A P, DE VET H C, DE BIE R A, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus[J]. Journal of Clinical Epidemiol, 1998, 51(12): 1235-1241.

[27] 王家良 . 临床流行病学：临床科研设计、测量与评价 [M]. 上海：上海科学技术出版社 , 2009.

[28] 夏惴惴 , 詹思延 . 如何撰写高质量的流行病学研究论文 第七讲 随机对照试验 Meta 分析的报告规范：QUOROM 介绍 [J]. 中华流行病学杂志 , 2007, 28(6): 618-620.

[29] MA L L, WANG Y Y, YANG Z H, et al.

doi:10.3772/j.issn.2095-915x.2025.03.004

Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? [J]. Military Medical Research, 2020, 7(1): 7.

[30] MOHER D, SCHULZ K F, ALTMAN D, et al. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials[J]. The Lancet, 2001, 285(15): 1987-1991.

[31] 王波 , 詹思延 . 如何撰写高质量的流行病学研究论文第三讲诊断试验准确性研究的报告规范 —— STARD 介绍 [J]. 中华流行病学杂志 , 2006, 27(10): 909-912.

[32] VON ELM E, ALTMAN D G, EGGER M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies[J]. The Lancet, 2007, 370(9596): 1453-1457.

[33] 罗晓敏 , 詹思延 . 如何撰写高质量的流行病学研究论文第六讲非随机对照试验研究报告规范 —— TREND 介绍 [J]. 中华流行病学杂志 , 2007, 28(4): 408-410.

[34] STERNE J A C, SAVOVIĆ J, PAGE M J, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials[J]. BMJ, 2019(366): 14898.

[35] STERNE J A C, HERNAN M A, REEVES B C, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions[J]. BMJ, 2016(355): i4919.

[36] WHITING P, RUTJES A W, REITSMA J B, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews[J]. BMC Med Research Methodology, 2003(3): 25.

[37] WELLS G, SHEA B, O'CONNELL D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses[EB/OL]. (2021-5-3) [2024-03-01]. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

[38] AROMATARIS E, FERNANDEZ R, GODFREY C M, et al. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach[J]. International Journal of Evidence-Based Healthcare, 2015, 13(3): 132-140.

[39] GUYATT G H, OXMAN A D, VIST G E, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations[J]. BMJ, 2008, 336(7650): 924-926.

[40] 栾嘉 , 邓强庭 , 李玥 , 等 . 医学临床研究论文质量评价指标体系构建 [J]. 中国科技期刊研究 , 2023, 34(2): 144-155.

[41] MCMASTER UNIVERSITY. McMaster health knowledge refinery[EB/OL]. (2024-03-01) [2024-03-01]. https://hiru.mcmaster.ca/hiru/hiru_hedges_home.aspx.

[42] WILCZYNSKI N L, MORGAN D, HAYNES R B, et al. An overview of the design and methods for retrieving high-quality studies for clinical care[J]. BMC Medical Informatics and Decision Making, 2005(5): 20.

[43] MCMASTER UNIVERSITY. McMaster health knowledge refinery-table 2[EB/OL]. (2024-03-01) [2024-03-01]. https://hiruweb.mcmaster.ca/hkr/hedges/table-2/.

[44] MCMASTER UNIVERSITY. McMaster health knowledge refinery-table 3[EB/OL]. (2024-03-01) [2024-03-01]. https://hiruweb.mcmaster.ca/hkr/hedges/table-3/.

[45] ACP JOURNAL CLUB. ACP journal club inclusion criteria[EB/OL]. (2012-11-14) [2024-03-01]. https://www.acpjournals.org/journal/aim/acpjc/inclusion-criteria.

[46] MOHER D, LIBERATI A, TETZLAFF J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement[J]. PLoS Medicine, 2009, 6(7): e1000097.

[47] BIAN J, MORID M A, JONNALAGADDA S, et al. Automatic identification of high impact articles in PubMed to support clinical decision making[J]. Journal of Biomedical Informatics, 2017(73): 95-103.

[48] COCHRANE. Cochrane database of systematic reviews[EB/OL]. (2024-03-01) [2024-03-01]. https://www.cochranelibrary.com/cdsr/about-cdsr.

[49] COCHRANE. Cochrane handbook for systematic reviews of interventions[EB/OL]. (2024-8-22) [2025-05-28]. https://training.cochrane.org/handbook/

current.

[50] WONG S S, WILCZYNSKI N L, HAYNES R B, et al. Developing optimal search strategies for detecting sound clinical prediction studies in MEDLINE[J]. AMIA Annual Symposium Proceedings, 2003: 728-732.

[51] GILL P J, ROBERTS N W, WANG K Y, et al. Development of a search filter for identifying studies completed in primary care[J]. Family Practice, 2014, 31(6): 739-745.

[52] NIH. Frequently Asked questions about indexing for MEDLINE[EB/OL]. (2023-02-24) [2024-03-01]. https://pubmed.ncbi.nlm.nih.gov/help/#clinical-study-category-filters.

[53] LEEFLANG M, SCHOLTEN R, RUTJES A, et al. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies[J]. Journal of Clinical Epidemiology, 2006, 59(3): 234-240.

[54] TSAFNAT G, GLASZIOU P, KARYSTIANIS G, et al. Automated screening of research studies for systematic reviews using study characteristics[J]. Systematic Reviews, 2018, 7(1): 64.

[55] TSAFNAT G, GLASZIOU P, KARYSTIANIS G, et al. Text summarization in the biomedical domain: a systematic review of recent research[J]. Journal of Biomedical Informatics, 2014(52): 457-467.

[56] DEL FIOL G, MICHELSON M, IORIO A, et al. A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: comparative analytic study[J]. Journal of Medical Internet Research, 2018, 20(6): e10281.

[57] AFZAL M, HUSSAIN M, HAYNES R B, et al. Context-aware grading of quality evidences for evidence-based decision-making[J]. Health Informatics Journal, 2019, 25(2): 429-445.

[58] MARSHALL I J, NOEL-STORR A, KUIPER J, et al. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide[J]. Research Synthesis Methods, 2018, 9(4): 602-614.

[59] APHINYANAPHONGS Y, TSAMARDINOS I, STATNIKOV A, et al. Text categorization models for high-quality article retrieval in internal medicine[J]. Journal of the American Medical Informatics Association, 2005, 12(2): 207-216.

[60] BIAN J, ABDELRAHMAN S, SHI J, et al. Automatic identification of recent high impact clinical articles in PubMed to support clinical decision making using time-agnostic features[J]. Journal of Biomedical Informatics, 2019(89): 1-10.

[61] LIN J, CHANG C, LIN M, et al. Automating the process of critical appraisal and assessing the strength of evidence with information extraction technology[J]. Journal of Evaluation in Clinical Practice, 2011, 17(4): 832-838.

[62] ABDELKADER W, NAVARRO T, PARRISH R, et al. A deep learning approach to refine the identification of high-quality clinical research articles from the biomedical literature: protocol for algorithm development and validation[J]. JMIR Research Protocols, 2021, 10(11): e29398.

[63] LOKKER C, BAGHERI E, ABDELKADER W, et al. Deep learning to refine the identification of high-quality clinical research articles from the biomedical literature: performance evaluation[J]. Journal of Biomedical Informatics, 2023(142): 104384.

[64] 毛宗福 , 丁元林 . 临床诊断试验论著质量分析 [J]. 中华医院管理杂志 , 1998(2): 42-43.

[65] 陈飞 , 曾艳彩 , 魏厚仁 . 从循证医学角度分析中华眼科杂志已发表临床疗效研究论文的质量 [J]. 中华眼科杂志 , 2004(9): 35-39.

[66] 尚淑贤 , 颜艳 , 吴晓初 . 中文临床医学期刊随机对照临床试验摘要报告质量评估 [J]. 中国科技期刊研究 , 2017, 28(12): 1121-1127.

[67] 赵宏杰 , 张俊华 , 郭利平 , 等 . 我国高影响因子中文医学期刊发表随机对照试验注册、伦理、知情同意报告现状及质量评价分析 [J]. 中国循证医学杂志 , 2018, 18(7): 735-739.