机器学习用于识别发热婴儿菌血症和脑膜炎：一项系统综述

摘 要

背景 准确识别有侵袭性细菌感染 (IBI) 风险的发热婴儿，如菌血症和脑膜炎，对于减少相关不良结局并避免不必要的侵入性操作至关重要。机器学习 (ML) 在提高其诊断准确性方面的作用，仍有待进一步评估。

目的 系统评价基于 ML 的模型在识别 90 d 以内发热婴儿 IBI 的诊断效能。

数据来源 系统检索 MEDLINE、Scopus、Embase、CINAHL、CENTRAL、ClinicalTrials.gov 以及世界卫生组织国际临床试验注册平台数据库，检索时间截至 2024 年 12 月。

研究选择 符合纳入标准的研究均将 ML 模型应用于 90 d 以内的发热婴儿以识别 IBI，并报告诊断性能指标，同时以培养确诊感染作为参考标准。在筛选的 4 756 条记录中，最终 6 项研究符合纳入标准。

数据提取 2 名研究者独立提取研究特征、模型类型、预测结局和准确性指标。采用改良的诊断准确性研究质量评价工具第 2 版 (QUADAS-2) 评估偏倚风险。由于研究之间存在显著的方法学异质性，因此无法进行 meta 分析。

结果 6 项研究评估了多种 ML 模型，包括 logistic 回归、随机森林、神经网络、支持向量机和集成方法。模型敏感度为 57% ～ 100%，特异度为 30% ～ 94%，ROC-AUC 范围为 0.57 ～ 0.9。多个模型在预测 IBI 方面优于传统工具（如 PECARN、Step-by-Step），尤其在特异度方面表现突出。

局限性 局限性包括多项研究采用回顾性数据收集方式、数据处理的报告有限以及缺乏外部验证。

结论 ML 模型在识别 IBI 风险婴儿方面显示出良好的效能，并且在提高特异度方面优于传统工具。未来研究应系统整合并透明报告模型可解释性分析，以提高模型的临床可解释性。在临床推广应用之前，必须进行更广泛的验证，推进方法学标准化，并进行谨慎的临床整合。

Giacomo Ferraris, MD,[1,a] Sara Tazzoli, MD,[1,a] Camilla Folisi, MD,[1] Laura Pisani, MD,[1] Hooman H. Rashidi, MD, MS, FCAP,[2,3] Nathan Kuppermann, MD, MPH,[4] Silvia Bressan, MD, PhD[1]

[1]Department of Women's and Children's Health, Padua University Hospital, Italy; [2]Computational Pathology & AI Center of Excellence (CPACE), the University of Pittsburgh, School of Medicine, Pittsburgh, Pennsylvania; [3]Department of Pathology, School of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania; and [4]Departments of Pediatrics and Emergency Medicine, GW School of Medicine and Health Sciences, and Children's National Hospital, Washington, District of Columbia

[a]Contributed equally as co-first authors.

American Academy of Pediatrics
DEDICATED TO THE HEALTH OF ALL CHILDREN™

The article(s) originally appeared in English in the 2026, Volume 158 of Pediatrics, published by the American Academy of Pediatrics (AAP), and are hereby authorized for translation into Chinese by Children's Hospital of Soochow University.

引　言

发热婴儿的评估和处理对急诊科 (ED) 的临床医师而言是一项重大挑战。低龄发热婴儿常无明确感染源。虽然大多数为良性病毒感染，但约 10% 的低龄发热婴儿存在尿路感染 (UTI)[1]，1% ～ 2% 存在侵袭性细菌感染 (IBI)[1-3]，包括细菌性脑膜炎和菌血症或脓毒症。未能及时识别细菌感染婴儿可能导致严重的甚至致命性的并发症 [4,5]。虽然 UTI 可通过尿液分析快速识别，但为帮助临床医师在急性诊疗环境中及时识别 IBI 婴儿以避免延迟诊断，研究者已开发出多种风险分层标准。这些标准综合了不同组合的临床发现和实验室指标，近期也纳入了降钙素原 (PCT) 这一指标 [6-8]。

近期以 PCT 为生物标志物的风险分层工具包括 Step-by-Step 规则和简化版儿科急诊应用研究网络 (PECARN) 发热婴儿风险评估规则 (表 1)。2 者对 IBI 的识别均表现出较高的敏感度 (分别为 92% 和 96.7%)[9,10]，其特异度虽高于未纳入 PCT 的旧模型，但仍不理想，约为 40% ～ 60%。较低的特异度导致更多的腰椎穿刺、住院治疗和经验性抗生素的使用 [9,11]。这些干预措施不仅对患儿家庭带成心理压力，还会引发潜在的不良反应和医源性风险，同时增加医院资源的使用，并且进一步加剧医院床位紧张 [9,10,12,13]。

近年来，机器学习 (ML) 算法在预测低龄发热婴儿中 IBI 风险方面显示出良好的应用前景，在保持较高敏感度的同时，提高了预测特异度。然而，ML 在输入变量如何转化为预测结果方面的透明度有限，加之各研究在方法学、算法和数据集方面的差异性，引发了对这些模型可解释性的担忧。这使得在 ED 对发热婴儿进行风险评估和临床决策时，难以将基于 ML 的预测结果与传统临床决策工具进行比较 [14,15]。

本系统综述旨在总结现有 ML 识别 IBI 高风险发热婴儿的研究证据。此外，本研究还旨在为临床医师和科研人员提供急性诊疗环境中基于 ML 预测低龄发热婴儿 IBI 风险的最新证据总结，重点关注这些模型的可解释性及其整合入临床工作流程的潜在价值。

方　法

本系统综述按照《Cochrane 干预性系统评价手册》6.3 版 [16] 进行，并依据《诊断试验准确性系统综述和 Meta 分析优先报告条目》(PRISMA-DTA) 声明 [17] 进行报告。研究方案已在国际系统综述前瞻性注册平台 (PROSPERO，网址：https://www.crd.york.ac.uk/PROSPERO/view/CRD42024521929) 注册。完整方案亦见补充附录。

检索策略

我们在以下医学电子数据库中进行检索：MEDLINE、EMBASE、Scopus、CINAHL 和 Cochrane 图书馆 [Cochrane 系统综述数据库、Cochrane 对照试验中央注册库 (CENTRAL)]，检索时间从建库起至 2024 年 12 月 31 日。同时，检索 ClinicalTrials.gov 和世界卫生组织国际临床试验注册平台。完整检索策略见补充附录。此外，我们对纳入研究的参考文献列表以及相关综述文献的参考文献进行手工检索，旨在发现其他可能的相关研究。仅纳入英文文献。

2 个评审组 (第 1 组：ST 和 GF；第 2 组：LP 和 CF) 在完成专门培训后，使用 Covidence 软件独立进行标题 - 摘要和全文筛选。筛选过程中若出现分歧，则通过与资深研究者 (SB) 进行结构化讨论以解决分歧，以确保筛选过程的一致性和透明性。对于全文筛选排除的文献，系统记录具体排除原因，以确保研究过程的透明性和可重复性。

纳入标准

纳入符合以下条件的回顾性和前瞻性观察性研究：应用 ML 预测 ED、急性诊疗机构或已住院的发热婴儿 (年龄 ＜ 90 d) 是否患有 IBI。虽然本综述的初

表 1　简化版 PECARN 发热婴儿规则与 Step-by-Step 规则及其准确性指标

规则特征	简化版 PECARN 规则	Step-by-Step 规则
适用人群	0 ～ 60 d 发热婴儿	0 ～ 90 d 发热婴儿
低危标准		
患儿外观	无危重表现	外观良好
尿液分析	无白细胞尿	无白细胞尿
ANC 水平	＜ 4000 / μL	＜ 10 000 / μL
PCT	＜ 0.5 ng / mL	＜ 0.5 ng / mL
CRP	未纳入	＜ 20 mg / L
敏感度	96.7% (83.3 ～ 99.4)	92% (84.3 ～ 96.0)
特异度	61.5% (59.2 ～ 63.9)	46.9% (44.8 ～ 49.0)
NPV	99.9% (99.4 ～ 100.0)	99.3% (98.5 ～ 99.7)

ANC：中性粒细胞绝对计数；CRP：C 反应蛋白；NPV：阴性预测值；PECARN：儿科急诊应用研究网络。

始研究范围包括应用 ML 预测更广泛的严重细菌感染（也包括 UTI）以及更具临床意义的 IBI 亚组（即菌血症或脓毒症和脑膜炎），但我们认识到 UTI（发热婴儿中最常见的严重细菌感染）通常可通过尿液分析快速识别，因此可能不需要基于 ML 的决策支持。因此，本系统综述最终聚焦于评估 ML 在预测发热婴儿 IBI 风险中的研究。

仅纳入经同行评议的已发表研究。排除叙述性综述、未经同行评议的文献、社论、评论和来信；关于非人源物种、成人、儿童、年龄超过 90 d 的婴儿或尚未从产房或新生儿重症监护病房出院的新生儿的研究；非英文文献；在联系通讯作者后仍无法获取全文的文献；以及仅报告 ML 预测更广泛严重细菌感染结果的文献。目标疾病定义为存在 IBI，即菌血症或脓毒症或细菌性脑膜炎。

纳入所有已发表的用于识别 IBI 婴儿的 ML 方法。这些方法包括监督学习、无监督学习和强化学习，通常分为左对齐或右对齐模型。左对齐模型以固定起始时点为基准，预测此后是否发生结局；右对齐模型则随时间动态更新输入数据，持续预测未来预定时间窗内的结局风险，即实时或连续预测。左右对齐模型均纳入本系统综述 [15]。

结局指标

本研究的主要结局指标是 ML 算法识别年龄 ≤ 90 d 发热婴儿中 IBI 的诊断准确性。次要结局：(1) 最佳 ML 算法与当前风险分层工具的比较（如适用于 ≤ 90 d 婴儿的 Step-by-Step[11]；适用于 ≤ 60 d 婴儿的 PECARN 发热婴儿预测规则 [9] 和 IBI 评分系统 [8]）；(2) 评估 ML 模型在以下亚组中的诊断准确性：新生儿（≤ 28 d）vs 婴儿（＞ 28 d），以及 ED 评估患儿 vs 已住院患儿。诊断试验性能通过以下指标评估：受试者工作特征曲线下面积 (ROC-AUC)、敏感度、特异度、阳性预测值 (PPV)、阴性预测值 (NPV)、阳性似然比 (LR+) 和阴性似然比 (LR-)。

数据提取

数据提取由 2 名研究者 (GF 和 ST) 独立完成。收集的信息包括研究特征，如第一作者姓名、发表年份和期刊、发表语言、国家、研究设计、参考标准、研究场所 (ED 或住院)、研究人群的具体信息（样本量、日龄、性别、纳入和排除标准、缺失值处理方法、结局事件发生患者数量）、研究结局定义 [SBI 和（或）IBI]、ML 算法（类型和模型特征、ML 算法训练和验证方法）以及模型性能评价指标（所提出算法的准确性评价指标）。

ML 模型研究的偏倚风险评估

我们使用修订版诊断准确性研究质量评估工具 (QUADAS-2)[18] 评估纳入研究的质量，该工具包括 4 个关键领域：患者选择、待评价试验、参考标准和流程与时间。考虑到 ML 研究的特殊性，我们对该工具进行了针对性调整。据我们所知，目前尚无广泛应用于人工智能 (AI) 诊断准确性研究的特定质量评价工具，这对 ML 模型评价的标准化构成挑战。在已发表的 AI 系统综述中，很大比例 (94%) 未使用 AI 特定质量评价工具 [19]。因此，目前最广泛使用的工具之一是 QUADAS-2，该工具可用于评价偏倚风险和适用性，并得到当前 PRISMA 2020 指南的推荐 [20]。目前，针对 AI 研究设计的 QUADAS-2 扩展工具（称为 QUADAS-

AI) 正在开发中[19]，但尚未用于研究评价。为解决这些局限性，我们提出了针对 AI 研究特点进行修订的 QUADAS-2 扩展版本，以更有效地评估相关研究中的偏倚风险[21,22]，详见补充表 1。2 名研究者独立进行评估 (CF 和 LP)。若有分歧，则通过讨论及第 3 名独立评审者 (SB) 的参与解决。

数据综合

由于纳入或排除标准、分析变量、研究结局和 ML 技术存在显著异质性，无法通过 meta 分析得出

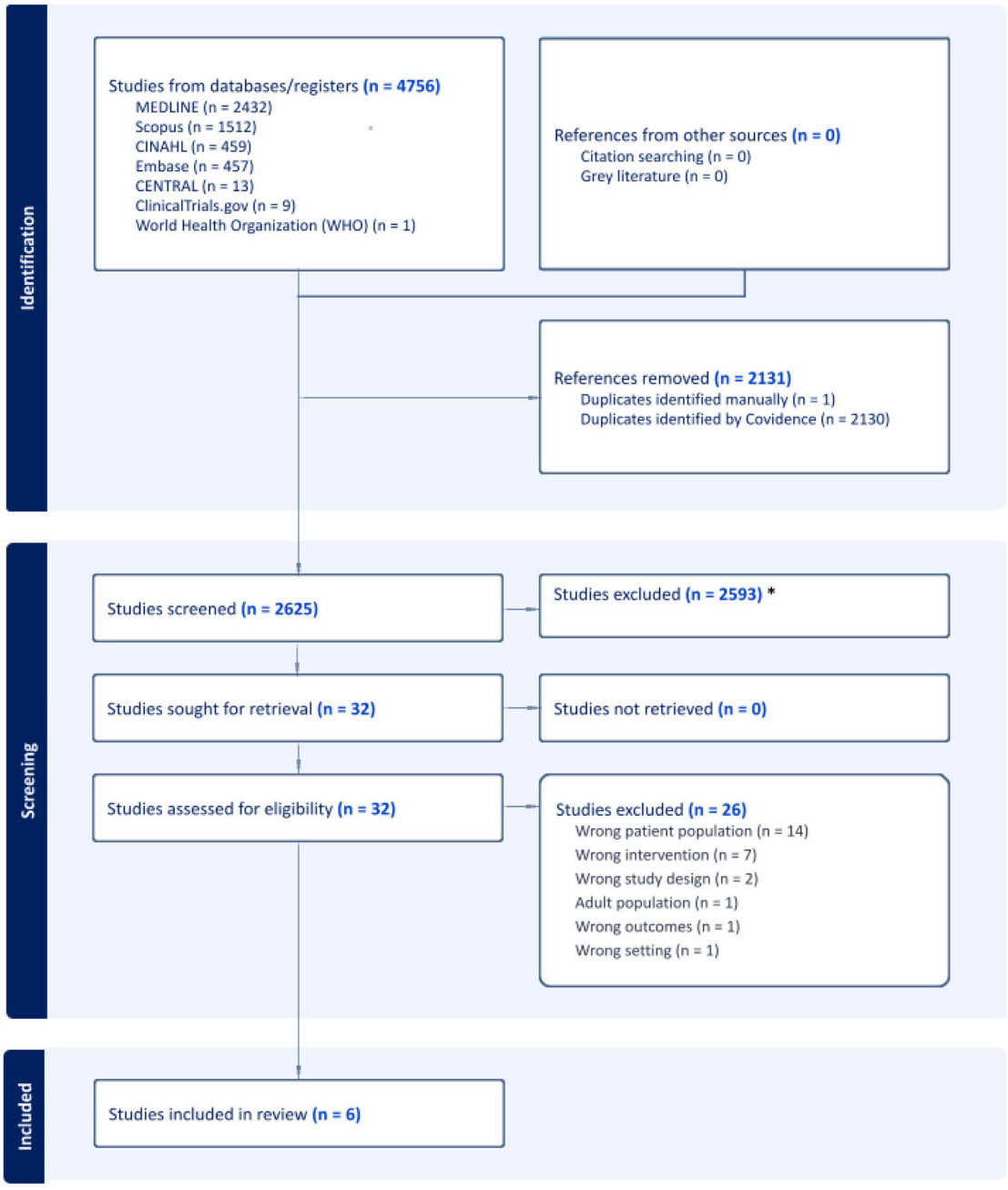


图 1　研究筛选过程流程

*研究对象不符合标准 (679 篇)、干预措施不符合标准 (446 篇)、研究设计不符合标准 (213 篇)、成人研究人群 (285 篇)、结局指标不符合标准 (368 篇)、研究场景不符合标准 (602 篇)。

有意义的诊断效能指标合并估计值。因此，我们采用叙述性综合方法，对研究结果进行定性分析。

结　果

总体而言，通过检索策略共获得 4 756 条记录。经 Covidence 软件去重后，2 625 条记录进入题名和摘要筛选；其中 32 篇全文文献进入资格评估，9 篇被认为符合纳入标准 [14,23-30]。经仔细审查，发现其中 3 篇研究仅将 ML 用于涵盖更广泛严重细菌感染的数据集，且未提供针对 IBI 的单独分析或结局数据，因此予以排除 [24,26,27]。所以，最终本系统综述共纳入 6 项

表 2　纳入研究的特征

研究	纳入标准	排除标准	数据收集	样本量 (N)	男性，n (%)	新生儿，n (%)	中位年龄 (IQR)
Ramgopal 等 [29]，2020	• 0 ～ 60 d • 发热（前 24 h 内体温＞ 38 ℃） • 就诊于 ED	• 早产 • 数据缺失（血、尿或脑脊液培养） • 既存疾病 • 临床脓毒症 • 近期抗生素使用 • 软组织感染 • 未行腰椎穿刺者在急诊就诊后 8 ～ 14 d 无电话或病历随访	前瞻性	1 470	859 (58)	456 (31)	37 d (25 ～ 48)
Chiu 等 [14]，2021	• 0 ～ 60 d • 发热（主观描述或急诊入院时记录） • 就诊于 ED	• 早产 • 数据缺失（血、尿或脑脊液培养） • 实验室数据缺失 • 复杂慢性疾病 • 从其他医院转诊	回顾性	4 211	2 462 (58)	未报告	32 d（未报告）
Yaeger 等 [25]，2022	• 0 ～ 90 d • 发热（急诊到达 6 h 内体温≥ 38° C） • 有记录胎龄 • 发热和病程持续时间 • 在 ED 进行初次评估	未明确	回顾性	2 311	1 220 (53)	473 (20)	54 d（未报告）
Yang 等 [30]，2023	• 0 ～ 60 d • 发热（在家、分诊或急诊入院时体温＞ 38 ℃）	• 非发热主诉 • 数据缺失（血、尿或脑脊液培养） • 复杂慢性疾病 • 违背医嘱出院 • 从其他医院转诊	回顾性	1 847	1 065 (58)	未报告	D：31 d (13 ～ 49) T：29 d (11 ～ 47)
Brouwer 等 [23]，2024	• 0 ～ 60 d • 就诊于 ED • 因疑似 SBI 接受检查 • 行 FBC 和 CRP 检测	• FBC、CRP 或血培养结果缺失	回顾性	2 692	1 515 (56)	未报告	34 d (21 ～ 45)
Ballard 等 [28]，2024	• 7 ～ 60 d； • 发热（急诊到达 4 h 内、医疗机构就诊前 12 h 内或急诊就诊时体温≥ 38 ℃）	• 在家发热但在医疗机构就诊前 12 h 内及急诊就诊 4 h 内无发热者 • 从其他医院转诊 • 同一日多次就诊者仅纳入首次	回顾性	4 411	2 558 (58)	未报告	39 d (24 ～ 51)

CRP：C 反应蛋白；CSF：脑脊液；D：诊断；ED：急诊科；FBC：全血细胞计数；LP：腰椎穿刺；T：测试。

研究。研究筛选过程的 PRISMA 流程见图 1。

5 项研究仅纳入年龄小于 60 d 的发热婴儿[14,23,28-30]，而 1 项研究将纳入标准扩展至 90 d 以内[25]。发热通常定义为在家庭、分诊或急诊就诊期间体温 ≥ 38° C；然而，其中 1 项研究纳入家长主观报告的发热[14]，另一项研究规定了精确的检测时间窗（如急诊到达 6 h 内[25]）。在纳入主观报告发热的研究中，作者未报告符合这些描述的患儿人数。这些婴儿与客观测量发热的患儿一起进行分析[14]。各研究具体纳入和排除标准详见表 2。

所有研究均在 ED 进行，大多数 (5 项）采用回顾性数据收集方法。研究样本量差异较大，范围为 1 470 ～ 4 411 例，这可能影响研究结果的可推广性，并提示需要谨慎处理数据。各研究中，男性婴儿比例为 53% ～ 58%。仅 2 项研究明确报告了新生儿（≤ 28 d) 的比例，范围为 20% ～ 31%。纳入婴儿的中位年龄为 29 ～ 54 d，各研究的 IQR 存在差异。研究结局的定义、总体发生率及按感染类型分类情况见表 3。各研究中发生结局事件的婴儿比例为 1.7% ～ 4.4%。

总体而言，6 项纳入研究共测试了 18 种 ML 模型。其中，集成方法 (EM) 的应用最为广泛，包括堆叠泛化 (1 项研究）、自助聚合法 (bagging) (3 项研究）和提升法 (boosting) (3 项研究）。广义线性模型 (GLM) 也被频繁使用，出现在 5 项研究中。支持向量机 (SVM) 应用于 3 项研究，神经网络应用于 2 项研究。基于实例的学习仅见于 1 项研究 (11%)。纳入研究中缺失数据处理方法和所用 ML 算法的详细信息见补充表 2。

各研究中最常纳入的变量包括年龄、最高体温和性别，这些变量几乎均被所有模型采用。实验室指标包括白细胞计数 (WBC)、中性粒细胞绝对计数 (ANC)、C 反应蛋白 (CRP) 和 PCT，在 5 项研究中以不同的组合形式予以使用（表 4)。纳入变量的分布见图 2。

各模型的关键统计指标，包括敏感度、特异度、PPV、NPV、LR+ 和 LR- 以及 ROC-AUC，见表 4。各算法

表 3　用于预测性 ML 模型训练和验证的研究结局

研究	结局	结局定义	发生结局事件的婴儿数，n (%)	按类型分类的 IBI 婴儿数，n (%)
Ramgopal 等[29]，2020	IBI	细菌性脑膜炎、菌血症；培养阳性	28 (1.9)	• 菌血症 20 (1.4)； • 细菌性脑膜炎 8 (0.5)
Chiu 等[14]，2021	IBI	细菌性脑膜炎、菌血症；培养阳性	126 (3.1)	• 菌血症 117 (2.8)； • 细菌性脑膜炎 18 (0.4)
Yaeger 等[25]，2022	IBI	细菌性脑膜炎、菌血症；培养阳性	39 (1.7)	• 菌血症或脑膜炎 39 (1.7)； • UTI 合并菌血症 15 (0.6)； • 仅菌血症 16 (0.7)； • 仅脑膜炎 3 (0.1)； • 菌血症合并脑膜炎 5 (0.2)
Yang 等[30]，2023	IBI	细菌性脑膜炎、菌血症；培养阳性	109 (2.8)	• 菌血症 48 (2.5)； • 细菌性脑膜炎 8 (0.4)
Brouwer 等[23]，2024	IBI	菌血症	76 (2.9)	• 不适用
Ballard 等[28]，2024	IBI	菌血症和（或）细菌性脑膜炎	196 (4.4)	• 菌血症 153 (4)； • 细菌性脑膜炎 43 (1)

IBI：侵袭性细菌感染；ML：机器学习；UTI：尿路感染。

表 4　纳入研究的结果

研究或算法	敏感度 (95% CI)	特异度 (95% CI)	PPV (95% CI)	NPV (95% CI)	LR+ (95% CI)	LR- (95% CI)	ROC-AUC (95% CI)	主要特征	误分类患者
Ramgopal 等[29]，2020									
随机森林	95.7 (78.1 ～ 99.9)	72.1 (69.8 ～ 74.4)	5.2 (3.3 ～ 7.7)	99.9 (99.5 ～ 100.0)	3.43 (3.05 ～ 3.87)	0.06 (0.01 ～ 0.41)	0.89 (0.83 ～ 0.95)	年龄、IBI 疑似度、体温、性别、WBC、ANC、尿液分析阳性、PCT	0
Chiu 等[14]，2021									
SVM	0.91 (0.07)	0.6 (0.03)	NA	NA	NA	NA	0.84 (0.03)	杆状核细胞、中性粒细胞、CRP	NA
XGBoost	0.9 (0.08)	0.57 (0.02)	NA	NA	NA	NA	0.84 (0.03)		
logistic 回归	0.9 (0.07)	0.59 (0.02)	NA	NA	NA	NA	0.85 (0.04)		
对照：Aronson IBI 评分	0.85 (0.06)	0.43 (0.01)	NA	NA	NA	NA	0.70 (0.03)		
Yaeger 等[25]，2022									
超级学习器							0.82 (0.76 ～ 0.87)		1 例脑膜炎
风险 = 0.01	1 (1 ～ 1)	0.03 (0.02 ～ 0.05)	0.02 (0.01 ～ 0.03)	1 (1 ～ 1)	1.04 (1.02 ～ 1.06)	0 (0 ～ 0)			
风险 = 0.03	0.97 (0.80 ～ 1)	0.53 (0.48 ～ 0.57)	0.03 (0.01 ～ 0.06)	0.99 (0.99 ～ 1)	2.07 (1.68 ～ 2.33)	0.05 (0 ～ 0.38)			
风险 = 0.05	0.85 (0.54 ～ 1)	0.70 (0.66 ～ 0.74)	0.05 (0.02 ～ 0.08)	0.99 (0.99 ～ 1)	2.83 (1.80 ～ 3.65)	0.22 (0 ～ 0.65)			
回归模型							0.8 (0.74 ～ 0.86)		1 例菌血症、1 例脑膜炎
风险 = 0.01	0.95 (0.75 ～ 1)	0.52 (0.47 ～ 0.56)	0.03 (0.01 ～ 0.06)	1 (0.99 ～ 1)	1.96 (1.52 ～ 2.24)	0.10 (0 ～ 0.40)			
风险 = 0.03	0.64 (0.29 ～ 1)	0.77 (0.74 ～ 0.81)	0.05 (0.01 ～ 0.09)	0.99 (0.98 ～ 1)	2.84 (1.27 ～ 4.41)	0.46 (0 ～ 0.95)			
风险 = 0.05	0.51 (0.17 ～ 0.86)	0.83 (0.79 ～ 0.86)	0.05 (0.01 ～ 1)	0.99 (0.98 ～ 1)	2.97 (0.94 ～ 5.20)	0.59 (0.17 ～ 1.01)			
Yang 等[30]，2023									
所有婴儿									
DLD	1.0	0.54	0.05	1.0	NA	NA	0.87	ANC、体温、心率、年龄、CRP	1 例菌血症
对照：Aronson IBI 评分	1.0	0.36	0.04	1.0	NA	NA	0.78		
对照：Step-by-Step 法则	1.0	0.35	0.04	1.0	NA	NA	0.67		
≤ 28 d 婴儿									
DLD	1.0	0.59	0.02	1.0	NA	NA	0.9		1 例菌血症
对照：Aronson IBI 评分	1.0	0.34	0.01	1.0	NA	NA	0.77		
对照：Step-by-Step 法则	1.0	0.15	0.01	1.0	NA	NA	0.57		
Brouwer 等[23]，2024									

子采样数据集 (n = 304)									
多因素 logistic 回归	0.57 (0.34 ～ 0.78)	0.94 (0.86 ～ 0.98)	0.75 (0.47 ～ 0.92)	0.88 (0.78 ～ 0.94)	NA	NA	0.82 (0.70 ～ 0.93)	年龄、CRP、E、L、N、PLT、NIr、MPVPR、PIr	NA
LDA	0.47 (0.25 ～ 0.70)	0.91 (0.82 ～ 0.96)	0.62 (0.35 ～ 0.84)	0.85 (0.75 ～ 0.92)	NA	NA	0.74 (0.61 ～ 0.87)		
K 近邻	0.52 (0.29 ～ 0.74)	0.91 (0.82 ～ 0.96)	0.64 (0.38 ～ 0.85)	0.86 (0.76 ～ 0.93)	NA	NA	0.76 (0.63 ～ 0.88)		
SVM	0.57 (0.34 ～ 0.78)	0.93 (0.84 ～ 0.97)	0.70 (0.44 ～ 0.89)	0.88 (0.78 ～ 0.94)	NA	NA	0.79 (0.68 ～ 0.91)		
随机森林	0.57 (0.34 ～ 0.78)	0.93 (0.84 ～ 0.97)	0.70 (0.44 ～ 0.89)	0.88 (0.78 ～ 0.94)	NA	NA	0.81 (0.69 ～ 0.92)		
决策树	0.61 (0.38 ～ 0.81)	0.85 (0.75 ～ 0.93)	0.56 (0.34 ～ 0.76)	0.88 (0.78 ～ 0.94)	NA	NA	0.72 (0.61 ～ 0.84)		
完整数据集 (n = 2 692)									
决策树	0.71 (0.59 ～ 0.80)	0.88 (0.87 ～ 0.89)	0.15 (0.11 ～ 0.19)	0.99 (0.98 ～ 0.99)	NA	NA	0.57 (0.55 ～ 0.59)		NA
随机森林	0.88 (0.78 ～ 0.94)	0.89 (0.88 ～ 0.90)	0.19 (0.15 ～ 0.24)	0.99 (0.99 ～ 0.99)	NA	NA	0.60 (0.58 ～ 0.62)		
2023 队列 (n = 197)									
决策树	0.57 (0.18 ～ 0.90)	0.52 (0.44 ～ 0.59)	0.04 (0.01 ～ 0.1)	0.97 (0.91 ～ 0.99)	NA	NA	0.58 (0.50 ～ 0.66)		NA
随机森林	0.71 (0.29 ～ 0.96)	0.62 (0.54 ～ 0.69)	0.06 (0.02 ～ 0.14)	0.98 (0.94 ～ 0.99)	NA	NA	0.59 (0.52 ～ 0.67)		
Ballard 等[28]，2024									
logistic 回归	NA	NA	NA	NA	NA	NA	0.831	WBC、体温、ANC、杆状核细胞、年龄、尿液分析、白细胞酯酶	NA
LASSO	NA	NA	NA	NA	NA	NA	0.837		NA
决策树	NA	NA	NA	NA	NA	NA	0.500		NA
随机森林	NA	NA	NA	NA	NA	NA	0.812		NA
XGBoost	NA	NA	NA	NA	NA	NA	0.839		NA

ANC：中性粒细胞绝对计数；BAND：杆状核细胞计数；CRP：C 反应蛋白；DLD：深度学习模型；E：嗜酸性粒细胞；HR：心率；IBI：侵袭性细菌感染；L：淋巴细胞；LASSO：最小绝对收缩和选择算子；LDA：线性判别分析；LE：白细胞酯酶；LR：logistic 回归；LR-：阴性似然比；LR+：阳性似然比；MLR：多因素 logistic 回归；MPVPR：平均血小板体积或血小板计数比值；N：中性粒细胞；NA：不适用；NLR：中性粒细胞或淋巴细胞比值；NPV：阴性预测值；PCT：降钙素原；PLR：血小板或淋巴细胞比值；PLT：血小板；PPV：阳性预测值；ROC-AUC：受试者工作特征曲线下面积；SE：敏感度；SP：特异度；SVM：支持向量机；UA：尿液分析；WBC：白细胞计数。

ROC-AUC 说明：ROC-AUC 是衡量检测方法或模型区分有无某种疾病患者（本研究中为有无 IBI）能力的指标，反映随机选取的 1 例有病患者的预测风险高于无病患者的可能性。 数值范围为 0.5（无判别能力）至 1.0（完全判别）。

敏感度范围为 57% ～ 100%，ML 模型和传统风险分层工具的值常超过 95%。与传统方法相比，部分 ML 模型表现出更高的特异度，尤其是随机森林和深度学习模型表现出优异的特异度，而部分模型的特异度较低，尤其是在采用较低风险阈值进行分类时更为明显。各算法的 ROC-AUC 值均显示出较高的区分能力，常超过 0.9。

最常用的预测因素包括临床参数（如年龄、体温）、血液检测指标（如 CRP、PCT、WBC、ANC）和尿液分析结果。其中，1 篇研究仅采用临床特征进行建模 [25]，而 5 项研究则同时纳入临床特征和实验室检测指标 [14,23,28-30]。ML 模型采用的特征类型以及模型预测所用的主要特征见图 2 和表 4。

关于模型可解释性，纳入研究中对解释性分析的报告并不一致（如基于树模型和集成模型的研究通常报告了全局特征重要性排序，而 GLM 则主要通过模型系数进行解释）；然而，现代可解释人工智能技术，如 SHAP 值或其他局部归因方法很少被应用。此外，研究中通常未系统评估这些解释结果是否符合临床实际意义。

关于次要结局指标，纳入研究中采用以下风险分层工具和预测模型作为所开发 ML 模型的比较标准：2 项研究使用 PECARN 发热婴儿规则，2 项使用 Step-by-Step 方法，3 项使用 IBI 评分，1 项使用 Rochester 低危标准（补充表 2 和表 4）。目前，没有研究针对次要结局中定义的年龄分层（新生儿 ≤ 28 d vs 婴儿 ＞ 28 d）专门开发并验证 ML 模型。然而，其中 1 项研究 [30] 在测试集中将 ML 模型应用于 ≤ 28 d 的亚组患儿。

质量评估

本系统综述纳入研究的偏倚风险见补充表 3。4 项研究 [23,28-30] 被评定为低偏倚风险，2 项被评定为高风险 [14,25]。研究被评为高偏倚风险的原因包括：用于评价研究结局（待评价试验）的样本量不足，未明确说明超参数调优方法，或缺乏在研究环境之外开展的验证。

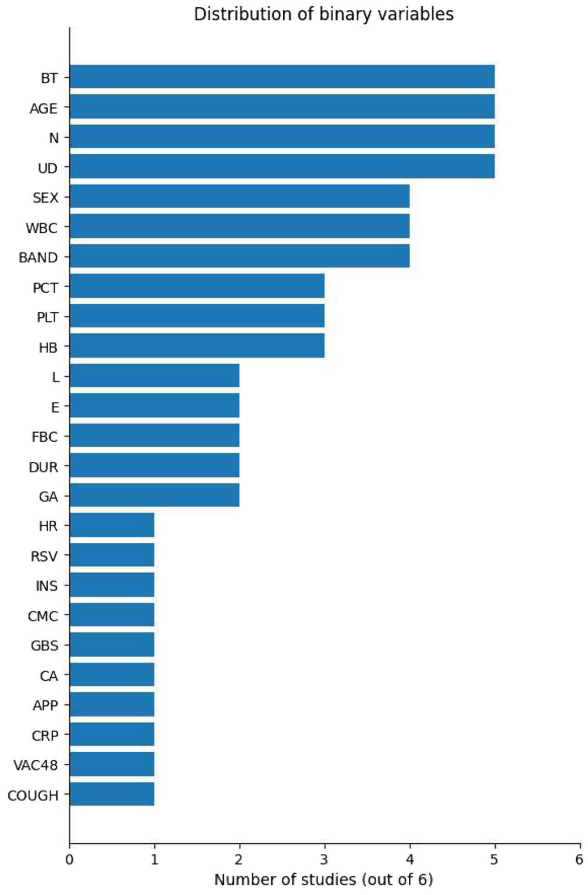


图 2　二元变量分布

APP：外观；BAND：杆状核细胞计数；BT：体温；CA：先天性异常；CMC：慢性疾病；CRP：C 反应蛋白；DUR：抗生素治疗持续时间；E：嗜酸性粒细胞；FBC：全血细胞计数；GA：胎龄；GBS：分娩住院期间 B 族链球菌感染相关抗生素使用；HB：血红蛋白；HR：心率；INS：保险类型；L：淋巴细胞；N：中性粒细胞；PCT：降钙素原；PLT：血小板；RSV：呼吸道合胞病毒感染状态；UD：尿液试纸检测；VAC48：48 h 内接种疫苗；WBC：白细胞。

讨　论

本系统综述评估了 ML 模型识别年龄 ≤ 90 d 发热婴儿 IBI 的诊断性能，发现多种 ML 方法均表现出良好的判别性能，其预测性能与现有成熟的临床风

险分层工具相当，甚至优于部分工具，包括 PECARN 规则、Step-by-Step 方法和 IBI 评分 [14,27,29,30]。我们的研究结果与既往研究一致，说明 ML 在感染性疾病结局预测中具有潜在的应用价值 [15]。

本综述中，ML 方法通常在具有相当的敏感度水平下，表现出更高的特异度，这支持其有潜力作为基于传统统计学方法建立的预测工具的辅助方法或替代方案。更高的特异度意味着在维持较高敏感度、尽量减少漏诊的前提下，有望减少不必要的腰椎穿刺、住院和抗生素治疗。考虑到不必要干预可能带来的药物不良反应、医源性并发症和对患儿家庭造成的心理压力 [4,5]，经过充分验证并适用于不同人群的准确 ML 模型，有望显著改善临床决策过程。

遵循既避免遗漏潜在危及生命的感染，又尽量减少不必要的侵入性检查、抗生素治疗和潜在医源性风险的原则，美国儿科学会 (AAP) 针对 8 ～ 60 d、一般状况良好的发热婴儿制定的临床实践指南，采用基于年龄分层的管理框架，以预先设定的炎症标志物阈值和结构化算法指导患儿检查策略和临床处理 [31]。在此背景下，ML 方法可提供互补性的优势。不同于传统规则算法主要依赖实验室指标的二分类截断值和临床表现的二分类分组，ML 模型可在无预设阈值的情况下纳入连续变量、复杂的非线性交互关系和概率性预测结果。此外，由于要积累足够大的样本，以准确预测细菌性脑膜炎等罕见结局，仍是研究面临的主要挑战，多中心 ML 模型开发依托大规模数据集，可能有助于进一步优化个体化风险评估。ML 模型并非用于取代现有算法，而是可作为辅助工具，在 AAP 指南 [31] 所阐述的基于年龄的结构化框架内，对风险评估进行进一步的精细化。

本系统综述还发现，纳入研究之间存在显著的异质性，主要体现在研究设计、研究人群、用于开发模型的临床及实验室特征，以及 ML 模型的开发和验证流程等方面。尽管 ML 模型总体表现良好，但并非所有研究均显示其相较于传统方法构建的模型具有一致的优势。尤其是基于小样本量或数据类别不平衡的数据集（如事件发生率低的单中心研究）构建的模型，在完整数据集应用时仅表现出中等的预测准确性 [23]。大多数研究纳入了临床和实验室预测指标，如年龄、最高体温、性别、CRP、PCT 和 ANC 等，这些指标在特征贡献度排序中始终位列前列；然而，用于确定并呈现这些特征重要性排序的可解释性方法在不同研究中的报告情况并不一致。1 项研究 [25] 仅依赖临床评估变量，未纳入实验室生物标志物，而后者在低龄婴儿 IBI 风险分层中研究最为广泛。尽管该研究开发的 ML 模型表现出良好的预测性能，但仅基于临床评估的模型在实践中难以标准化，因为不同临床医师在管理低龄发热婴儿方面的经验、能力和临床处置信心存在异质性。此外，仅少数开发的 ML 模型进行了内部或外部验证 [14,26,28,29]，凸显了标准化验证流程以确保模型可靠性和可推广性的重要性 [32,33]。这些发现提示，未来在 ML 模型预测发热婴儿 IBI 的研究中仍有进一步的改进空间，从而优化研究设计并提高方法学质量。另一方面，纳入研究中 IBI 的患病率，包括菌血症 (1.4% ～ 4%) 和脑膜炎 (0.3% ～ 1%) 与已报告的发热婴儿真实世界估计值大体一致，这支持了研究人群具备总体代表性 [3-6,8-11,31]；然而，事件发生率的差异和结局的绝对数量较低仍可能影响模型性能，尤其在校准度和类别不平衡方面。

输入转化为预测过程中的透明度有限 —— 通常被称为许多 ML 模型（特别是神经网络和集成算法）的“黑箱”性质 —— 传统上引发了人们对可解释性和临床信任的担忧。然而，透明度与信任之间的关系正在发生演变。临床医师越来越多地使用并依赖内部机制并非完全可解释的 AI 驱动工具，这表明信任度可能不仅源于机制透明度，还源于确凿的技术性能、严谨的验证流程、良好的易用性以及与临床工作流程的契合度 [34]。

尽管如此，在发热婴儿 IBI 评估等高风险临床场

景中，模型的可解释性仍是一个关键的考量因素。除了区分能力指标以外，临床医师还需要了解哪些变量驱动了模型预测结果，不确定性如何被呈现，以及模型预测输出是否被有意义地融入决策流程中。因此，未来研究应在预测性能与可解释性策略之间取得平衡，如采用特征归因方法、校准度评估以及开展前瞻性实施研究评价模型对临床实践的实际影响，而不仅仅关注算法本身的准确性[35]。

尽管 ML 模型在提高诊断精度方面具有前景，但其整合入临床工作流程仍需谨慎考量。相比需要大量或非标准输入的模型，依赖就诊时常规可得变量（如生命体征和标准实验室参数）的模型在实际应用中往往更具可行性。将 ML 预测结果与经过验证的风险评分相结合，或将其作为决策支持工具整合入电子健康记录系统，可能成为模型预测性能和临床实际应用之间的有效桥梁。与电子病历的无缝整合、自动化数据提取以及顺应临床决策时效快速生成风险评估结果，对于避免增加临床医师的认知负担或延误诊疗至关重要。此外，还需开展评估可用性、临床医师接受度、决策的影响及患者结局等方面的前瞻性实施研究，以确定具有良好预测性能的 ML 模型是否能转化为有意义的临床获益[35-37]。ML 动态分层风险的潜力 —— 如针对新生儿与较大婴儿调整风险阈值，或整合特定临床情境下的相关参数 —— 也可能实现更加个体化的诊疗决策，从而有望减少不必要的腰椎穿刺、住院治疗和抗生素暴露。

本综述存在若干局限性。首先，研究之间的方法学异质性导致无法开展定量合并分析。其次，许多研究为回顾性，增加了因数据缺失或数据错误分类所导致的偏倚风险。第三，部分研究因数据处理和超参数优化方法的报告不充分，存在偏倚风险，从而限制了研究结果的可重复性。第四，所有纳入研究均未在不同临床环境中进行外部验证，限制了模型的推广应用能力。第五，仅纳入英文研究可能限制了国际代表性，进而影响了本研究结论的外部效度。最后，现有证据基础较为有限，且受到基础数据集地理范围和时间跨度的限制。纳入研究的数据大多来自单一国家或有限区域的研究队列，可能无法反映流行病学、检测实践或临床管理随时间发生的变化。

尽管存在这些局限性，除了总结当前的证据外，我们的研究结果对未来该领域 ML 模型的开发和实施具有重要意义。首先，研究设计、结局定义、预测变量选择和验证策略方面存在的显著异质性，凸显了进一步加强方法学标准化的必要性，包括对可解释性方法的规范报告。未来研究应优先开展前瞻性数据收集，透明报告数据预处理过程和超参数优化方法，并遵循预测模型研究的既定报告指南。其次，缺乏在不同医疗环境中的外部验证，引发了对模型可迁移性和真实世界适用性的担忧。鉴于临床工作流程、实验室检查可行性和患者人群的差异性，未来的 ML 模型应使用反映异质性实践环境的多中心数据集进行开发和验证。最后，鉴于 IBI 患病率较低但临床影响较大，模型开发应审慎处理类别不平衡、校准度和临床效用问题，确保预测性能能够转化为有意义的决策支持，而不仅仅体现为统计学上的判别能力。

结　论

ML 模型在识别伴有 IBI 的发热婴儿方面展现出良好的应用前景，其预测性能与基于传统统计学方法建立的临床预测规则相当，甚至可能优于后者。尽管这些发现令人鼓舞，但在 ML 模型广泛临床应用之前，仍需开展进一步的验证，加强模型开发过程的透明性，以及实现与临床诊疗路径的有效整合。未来研究应解决现有证据的局限性，重点关注多中心验证、前瞻性临床影响评估和以临床医师为中心的设计，以充分发挥 ML 在儿科急救医疗中的潜力。

缩略语

AI：人工智能

ANC：中性粒细胞绝对计数

CRP：C 反应蛋白

ED：急诊科

IBI：侵袭性细菌感染

LR-：阴性似然比

LR+：阳性似然比

ML：机器学习

NPV：阴性预测值

PCT：降钙素原

PECARN：儿科急诊应用研究网络

PPV：阳性预测值

QUADAS-2：诊断准确性研究质量评估工具 2

ROC-AUC：受试者工作特征曲线下面积

UTI：尿路感染

WBC：白细胞计数

（杨小凤 译　丁欣 校）

联系人：丁欣 etyxcbzx@suda.edu.cn

参考文献

1. Aronson PL, Thurm C, Alpern ER, et al; Febrile Young Infant Research Collaborative. Variation in care of the febrile young infant <90 days in US pediatric emergency departments. Pediatrics. 2014;134(4):667–677. PubMed doi:10.1542/peds.2014-1382
2. Simões e Silva AC, Oliveira EA. Update on the approach of urinary tract infection in childhood. J Pediatr (Rio J). 2015;91(6)(suppl 1):S2–S10. PubMed doi:10.1016/j.jped.2015.05.003
3. Biondi E, Evans R, Mischler M, et al. Epidemiology of bacteremia in febrile infants in the United States. Pediatrics. 2013;132(6):990–996. PubMed doi:10.1542/peds.2013-1759
4. Woll C, Neuman MI, Aronson PL. Management of the febrile young infant: update for the 21st century. Pediatr Emerg Care. 2017;33(11):748–753. PubMed doi:10.1097/PEC.0000000000001303
5. DePorre AG, Aronson PL, McCulloh RJ. Facing the ongoing challenge of the febrile young infant. Crit Care. 2017;21(1):68. PubMed doi:10.1186/s13054-017-1646-9
6. Milcent K, Faesch S, Gras-Le Guen C, et al. Use of procalcitonin assays to predict serious bacterial infection in young febrile infants. JAMA Pediatr. 2016;170(1):62–69. doi:10.1001/jamapediatrics.2015.3210
7. Dagan R, Powell KR, Hall CB, Menegus MA. Identification of infants unlikely to have serious bacterial infection although hospitalized for suspected sepsis. J Pediatr. 1985;107(6):855–860. PubMed doi:10.1016/S0022-3476(85)80175-X
8. Aronson PL, Wang ME, Shapiro ED, et al; Febrile Young Infant Research Collaborative. Risk stratification of febrile infants ≤60 days old without routine lumbar puncture. Pediatrics. 2018;142(6):e20181879. PubMed doi:10.1542/peds.2018-1879
9. Kuppermann N, Dayan PS, Levine DA, et al; Febrile Infant Working Group of the Pediatric Emergency Care Applied Research Network (PECARN). A clinical prediction rule to identify febrile infants 60 days and younger at low risk for serious bacterial infections. JAMA Pediatr. 2019;173(4):342–351. doi:10.1001/jamapediatrics.2018.5501
10. Sutiman N, Khoo ZX, Ong GYK, Piragasam R, Chong SL. Validation and comparison of the PECARN rule, Step-by-Step approach and Lab-score for predicting serious and invasive bacterial infections in young febrile infants. Ann Acad Med Singap. 2022;51(10):595–604. PubMed doi:10.47102/annals-acadmedsg.2022193
11. Gomez B, Mintegi S, Bressan S, Da Dalt L, Gervaix A, Lacroix L; European Group for Validation of the Step-by-Step Approach. Validation of the "Step-by-Step" approach in the management of young febrile infants. Pediatrics. 2016;138(2):e20154381. PubMed doi:10.1542/peds.2015-4381
12. Hu L, Shi Q, Shi M, Liu R, Wang C. Diagnostic value of PCT and CRP for detecting serious bacterial infections in patients with fever of unknown origin: a systematic review and meta-analysis. Appl Immunohistochem Mol Morphol. 2017;25(8):e61–e69. PubMed doi:10.1097/PAI.0000000000000552
13. Mintegi S, Bressan S, Gomez B, et al. Accuracy of a sequential approach to identify young febrile infants at low risk for invasive bacterial infection. Emerg Med J. 2014;31(e1):e19–e24. PubMed doi:10.1136/emermed-2013-202449
14. Chiu IM, Cheng CY, Zeng WH, Huang YH, Lin CR. Using machine learning to predict invasive bacterial infections in young febrile infants visiting the emergency department. J Clin Med. 2021;10(9):1875. PubMed doi:10.3390/jcm10091875
15. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. 2020;46(3):383–400. PubMed doi:10.1007/s00134-019-05872-y
16. Higgins JPT, Thomas J, Chandler J, et al, eds. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5. Cochrane; 2024. Accessed May 31, 2026. https://training.cochrane.org/handbook

17. McInnes MDF, Moher D, Thombs BD, et al; and the PRISMA-DTA Group. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMADTA Statement. JAMA. 2018;319(4):388–396. PubMed doi:10.1001/jama.2017.19163
18. Whiting PF, Rutjes AW, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–536. PubMed doi:10.7326/0003-4819-155-8-201110180-00009
19. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med. 2021;27(10):1663–1665. PubMed doi:10.1038/s41591-021-01517-0
20. Hanna MG, Pantanowitz L, Jackson B, et al. Ethical and bias considerations in artificial intelligence/machine learning. Mod Pathol. 2025;38(3):100686. PubMed doi:10.1016/j.modpat.2024.100686
21. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. Front Med. 2021;8:607952. doi:10.3389/fmed.2021.607952
22. Jayakumar S, Sounderajah V, Normahani P, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. NPJ Digit Med. 2022;5(1):11. PubMed doi:10.1038/s41746-021-00544-y
23. Brouwer L, Cunney R, Drew RJ. Predicting community acquired bloodstream infection in infants using full blood count parameters and C-reactive protein; a machine learning study. Eur J Pediatr. 2024;183(7):2983–2993. PubMed doi:10.1007/s00431-024-05441-6
24. Yaeger JP, Jones J, Ertefaie A, Caserta MT, van Wijngaarden E, Fiscella K. Using clinical history factors to identify bacterial infections in young febrile infants. J Pediatr. 2021;232:192–199.e2. PubMed doi:10.1016/j.jpeds.2020.12.079
25. Yaeger JP, Jones J, Ertefaie A, Caserta MT, Fiscella KA. Derivation of a clinical-based model to detect invasive bacterial infections in febrile infants. J Hosp Med. 2022;17(11):893–900. PubMed doi:10.1002/jhm.12956
26. Yaeger JP, Jones J, Ertefaie A, Caserta MT, van Wijngaarden E, Fiscella K. Refinement and validation of a clinical-based approach to evaluate young febrile infants. Hosp Pediatr. 2022;12(4):399–407. PubMed doi:10.1542/hpeds.2021-006214
27. Chong SL, Niu C, Ong GY, et al. Febrile infants risk score at triage (FIRST) for the early identification of serious bacterial infections. Sci Rep. 2023;13(1):15845. PubMed doi:10.1038/s41598-023-42854-z
28. Ballard DW, Huang J, Sharp AL, et al. An all-inclusive model for predicting invasive bacterial infection in febrile infants age 7–60 days. Pediatr Res. 2024;96(3):759–765. PubMed doi:10.1038/s41390-024-03141-3
29. Ramgopal S, Horvat CM, Yanamala N, Alpern ER. Machine learning to predict serious bacterial infections in young febrile infants. Pediatrics. 2020;146(3):e20194096. PubMed doi:10.1542/peds.2019-4096
30. Yang Y, Wang YM, Lin CR, et al. Explainable deep learning model to predict invasive bacterial infection in febrile young infants: A retrospective study. Int J Med Inform. 2023;172:105007. PubMed doi:10.1016/j.ijmedinf.2023.105007
31. Pantell RH, Roberts KB, Adams WG, et al; Subcommittee on Febrile Infants. Evaluation and management of well-appearing febrile infants 8 to 60 days old. Pediatrics. 2021;148(2):e2021052228. PubMed doi:10.1542/peds.2021-052228
32. Pantanowitz L, Pearce T, Abukhiran I, et al. Nongenerative artificial intelligence in medicine: advancements and applications in supervised and unsupervised machine learning. Mod Pathol. 2025;38(3):100680. PubMed doi:10.1016/j.modpat.2024.100680
33. Rashidi HH, Albahra S, Robertson S, Tran NK, Hu B. Common statistical concepts in the supervised machine learning arena. Front Oncol. 2023;13:1130229. PubMed doi:10.3389/fonc.2023.1130229
34. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. PubMed doi:10.1186/s12916-019-1426-2
35. Khan SD, Hoodbhoy Z, Raja MHR, et al. Frameworks for procurement, integration, monitoring, and evaluation of artificial intelligence tools in clinical settings: a systematic review. PLOS Digit Health. 2024;3(5):e0000514. PubMed doi:10.1371/journal.pdig.0000514
36. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337–1340. PubMed doi:10.1038/s41591-019-0548-6
37. Di Sarno L, Caroselli A, Tonin G, et al. Artificial intelligence in pediatric emergency medicine: applications, challenges, and future perspectives. Biomedicines. 2024;12(6):1220. PubMed doi:10.3390/biomedicines12061220

【英文原件请参阅 *PEDIATRICS 2026;158(2):e2025074576*】