Advances in Clinical Medicine 临床医学进展, 2025, 15(3), 1501-1506
Published Online March 2025 in Hans. https://www.hanspub.org/journal/acm
https://doi.org/10.12677/acm.2025.153769

Hans汉斯

非结构化数据在儿童脓毒症早期预测中的应用价值及进展

宋婉沁[1]，王浩林[2]，李　静[1*]

[1]重庆医科大学附属儿童医院重症医学科，国家儿童健康与疾病临床医学研究中心，儿童发育疾病研究教育部重点实验室，重庆市重点实验室，重庆
[2]重庆医科大学信息学院，重庆

收稿日期：2025年2月17日；录用日期：2025年3月9日；发布日期：2025年3月17日

摘　要

儿童脓毒症是感染诱发的可导致器官功能障碍和死亡的急危重症，早期识别和诊断对于防治脓毒性休克和改善预后至关重要。根据临床标准，医生需要结合患儿的临床表现、病史、体格检查和实验室检查等综合判断并诊断。随着医疗信息化的不断发展，储存在电子病历中的数据被得到更多的关注和挖掘，除了结构化数据外，非结构化数据约占总数据的80%，有着更为丰富的信息，近年来逐渐开始运用于各种疾病的诊疗预测模型中，且在早期诊断、个性化治疗方案制定及预后评估等方面均取得了显著性的进展。本文旨在对非结构化数据在脓毒症早期预测模型中的运用前景及其挑战进行综述，为儿童脓毒症的早期诊断、个性化治疗和预后方面研究提供理论基础。

关键词

脓毒症，非结构化数据，早期诊断，预测模型

Application Value and Progress of Unstructured Data in Early Prediction of Pediatric Sepsis

Wanqin Song[1], Haolin Wang[2], Jing Li[1*]

[1]Department of Intensive Care Unit, Children's Hospital of Chongqing Medical University, National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, Chongqing Key Laboratory, Chongqing
[2]School of Medical Informatics, Chongqing Medical University, Chongqing

[*]通讯作者。

文章引用: 宋婉沁, 王浩林, 李静. 非结构化数据在儿童脓毒症早期预测中的应用价值及进展[J]. 临床医学进展, 2025, 15(3): 1501-1506. DOI: 10.12677/acm.2025.153769

Received: Feb. 17th, 2025; accepted: Mar. 9th, 2025; published: Mar. 17th, 2025

Abstract

Pediatric sepsis is an acute and critical illness induced by infection that can lead to organ dysfunction and death. Early recognition and diagnosis are crucial for preventing septic shock and improving prognosis. According to clinical standards, doctors need to comprehensively assess and diagnose pediatric sepsis by considering the child's clinical manifestations, medical history, physical examination, and laboratory tests. With the continuous development of artificial intelligence, data stored in electronic health records (EHRs) has received increasing attention and exploration. In addition to structured data, unstructured data, which accounts for approximately 80% of the total data, contains richer information and has gradually been applied in predictive models for the diagnosis and treatment of various diseases in recent years. Unstructured data predictive models have increasingly become a research focus in the field of sepsis, achieving significant progress in early diagnosis, personalized treatment planning, and prognosis assessment. This article aims to review the application prospects and challenges of unstructured data in early predictive models for pediatric sepsis, providing a theoretical basis for research on early diagnosis, personalized treatment, and prognosis in pediatric sepsis.

Keywords

Sepsis, Unstructured-Data, Early Diagnosis, Predictive-Model

1. 引言

脓毒症是全球主要死亡原因之一，根据流行病学的研究，2017 年全球脓毒症年龄标准化死亡率为每 10 万人 148.1 例[1] [2]，其中儿童脓毒症的患病人数约为 2500 万，且死亡超过 300 万，是 5 岁以下儿童死亡的主要原因[3]。脓毒症的高死亡率及高费用负担是全球卫生系统面临的巨大挑战。随着对儿童脓毒症认知的深化，2024 年儿童国际组织基于 2016 年成人脓毒症诊断标准，提出了儿童脓毒症新的诊断评分标准——菲尼克斯脓毒症评分(Phoenix Sepsis Score, PSS)。该标准规定，评分 ≥2 分可确诊脓毒症；若心血管系统评分 ≥1 分，则诊断为脓毒性休克(排除了早产儿及围生期住院新生儿)。然而，菲尼克斯评分仅包含了呼吸、心血管、神经和凝血功能四个方面，此评分标准是否足够具有代表性还需在临床实践中进一步检验[4]。在治疗方面，抗生素的使用是脓毒症最关键的早期治疗方式之一[5]，有研究表明，抗菌药物延迟使用是脓毒症死亡率升高的一个关键因素[6]，故尽早识别或及时预测脓毒症在临床实践中具有重要意义。

近年来，在医学领域，人工智能已经被应用于药物发现、个性化诊断和治疗、分子生物学、生物信息学和医学成像等各领域。随着电子病历系统无纸化的发展，利用电子健康记录(Electronic Health Record, EHR)中丰富的临床数据对疾病的诊断、治疗及预后提供帮助成为可能[7]。目前已经能够通过应用程序提取和分析存储在电子医疗记录中的大量数据，并通过机器学习建模来早期识别并及时诊断脓毒症。有研究证明利用临床大数据及机器学习预测脓毒症的预测效能在一定程度上优于序贯器官衰竭评分

(Sequential Organ Failure Assessment, SOFA)、序贯器官衰竭评估快速评分(qSOFA)及全身炎症反应综合征(Systemic Inflammatory Response Syndrome, SIRS)评分[8] [9]等临床常用评分。目前多数研究聚焦于利用EHR中的结构化数据构建预测模型，然而，非结构化数据因其蕴含丰富信息，被视为未来疾病模型研究与构建的重要潜力开发领域。此外，当前研究大多聚焦于成人领域，而针对儿童脓毒症的预测模型相对匮乏，相应的临床证据也明显不足。本文旨在描述非结构化数据的数据处理方法，以及结合国内外研究现状，分析非结构化数据在儿童脓毒症预测模型中的潜能。

2. 非结构化数据的概念及处理

2.1. 数据的分类

电子医疗记录包含了以患者为主体的较为完整的医疗就诊信息，是重要的医疗数据资源，其在临床实践中有许多用处，可以更好地管理患者的医疗记录，提高护理质量，提供个性化的治疗，有效的预后评估[10]。

电子医疗记录的数据格式可分为两部分：结构化数据和非结构化数据。结构化数据包括患者人口统计数据、生命体征、实验室结果、药物等，约占整体数据的20%，剩下的80%为非结构化数据，是储存在电子病历中的叙述性数据，包括临床文本(如入院记录，病程记录、出院记录)、图文和影像等重要信息[11]。

2.2. 非结构化数据的处理与分析

结构化数据因其固有的结构特性，通常可通过标准统计或机器学习方法轻松分析。相比之下，非结构化信息(例如临床文本记录、护理记录等)难以被计算机直接识别和应用，故需先转换为计算机可解读的格式，方能进行后续的机器学习处理。传统方法常常依赖大量的人工工作，随着人工智能和深度学习的发展与进步，如今我们可以利用更为先进、有效的方式对非结构化数据进行处理，再加以利用[12] [13]。

研究初期，最常使用自然语言处理系统(Natural Language Processing, NLP)技术对非结构化数据进行处理，包括了常见的停用词去除、词元和词干分解、词频－逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)，词性标注(Part-of-Speech Tagging, POS Tagging)、命名实体识别(Named Entity Recognition, NER)、N-gram模型或潜在狄利克雷分配(LDA)主题建模，用于从文本中提取有用的特征，作为预测因子，进行预测模型的建立。2022年Goh等人[14]在脓毒症早期预测模型中使用LDA主题建模的方法对非结构化数据进行主题提取，并对其进行数值加权后与结构化数据相结合作为预测因子开发出一种新的SERA算法，并对其进行了测试，结果显示该模型可提前48小时预测脓毒症，并在脓毒症发生前12小时达到最高的预测准确性(AUC 0.94，敏感性0.87和特异性0.87)，并降低假阳性概率，同样使用主题建模处理非结构化数据进行脓毒症预测的还有Horng等人[15]的研究，也显示出不错的预测效能，训练数据集的主题模型AUC为0.86，验证数据集为0.86，测试数据集为0.85 (95% CI: 0.84~0.86)。

而近年来，随着深度机器学习的不断发展，神经网络(NN)显示出了较高的预测性能，使用神经网络嵌入(Embedding)已经成为研究热点。嵌入技术包括连续词袋(CBOW)模型、全局向量(GloVe)、Word 2Vec、BERT、Clinical BERT等不同方法，并逐渐在不同研究中显示出优势。其中，BERT是2018年开发出来的一种将单向模式改进为双向编码的程序，它优化了文档上下文嵌入模型的预测能力，并开发出了专门应用于临床叙述文本的ClinicalBERT模型[16]，在医学研究中得以应用。例如，2021年Amrollahi等人[17]使用ClinicalBERT嵌入的方法将非结构化数据进行转化，再与结构化数据相结合并形成最终向量，输入到预先训练的神经语言模型进行脓毒症预测，预测结果与仅用结构化数据相比，AUC从0.81提高到0.84，并减少了假阳性。相较于传统的NLP技术，神经网络嵌入技术等深度学习方法等能够更加深入地

挖掘文本中的语意信息，提高特征提取的准确性和全面性。然而，不同的嵌入方法也有各自的优缺点。例如，例如，CBOW 模型虽然计算速度快，但可能忽略上下文信息；GloVe 则考虑了全局词频信息，但在处理罕见词时可能效果不佳；Word2Vec 能够捕捉词语之间的相似度关系，但对上下文信息的利用仍有限；而 BERT 和 ClinicalBERT 虽然能够充分利用上下文信息，但计算复杂度和资源消耗相对较高。因此，在选择非结构化数据处理方法时，需要根据具体任务和数据特点进行综合考虑和评估。

3. 脓毒症早期预测模型的发展

基于大数据及机器学习的脓毒症预测模型通过分析患者的基本信息、病史数据、生物标志物、基因数据等多维度信息，可提前识别出潜在的脓毒症患者，提供及时的治疗建议和预后评估，对于脓毒症的未来诊疗具有极其重要的意义。

3.1. 成人脓毒症早期预测模型

3.1.1. 利用结构化数据的脓毒症早期预测模型

以往多数研究中，脓毒症诊断预测模型主要基于生命体征、实验室结果等结构化数据开发各类预测算法。在研究初期主要是基于生命体征进行模型建立。如 Barton 等人[18]利用心率、呼吸频率、体温、收缩压、舒张压和血氧饱和度(SpO_2)六种生命体征作为预测因子，开发了一种 MLA 算法。此算法可将脓毒症的预测时间提前 48 小时，模型 AUROC 为 0.83，表现出了良好的性能。同样，Mao 等人[19]利用 6 种生命体征开发出了另一种 InSight 算法，是第一个仅使用生命体征数据使得 AUROC 超过 0.9 的脓毒症预测模型。然而，若仅依据单一时间点的生命体征数据进行预测，则难以全面且实时地反映患者病情的动态变化，于是 2019 年，Wyk 等人[20]利用床旁监护仪收集的更为丰富且实时高频的生理数据(每分钟数据)，开发出了一种多层模型(以固定时间窗长度的方式作为数据层依次输入)算法。该模型定义符合 SIRS 标准至少两项为脓毒症发作，并利用脓毒症发作前 6 小时的数据进行预测，结果显示新模型可提前 204.87 ± 7.90 分钟预测脓毒症(相较于 SIRS 标准诊断时间)，并表现出了较好的性能(AUROC 为 0.79)。

随着实验室技术的进步，在生命体征的基础之上，加入蕴含更为丰富临床信息的实验室检验及检查指标以建立模型。2021 年，Wang 等人[21]利用国内某 ICU 的病人临床数据(收集每位患者 55 个变量，包括人口特征、生命体征、血常规、肝肾功能等与脓毒症发生密切相关指标)，并进行数据处理，最后筛选出 20 个特征性指标，利用随机森林算法进行模型建立，得到 AUROC 为 0.91 的模型，敏感性为 87%，特异性为 89%，有良好的识别能力。

3.1.2. 非结构数据在脓毒症早期预测模型中的应用

近几年，不断有研究将非结构化数据和结构化数据相结合来开发脓毒症诊断模型，更大程度地挖掘 EHR 中有价值的信息，从而提高了脓毒症诊断的准确性，提供及时治疗以改善预后[11]。

常用来进行模型预测的非结构化数据为临床文本，而临床文本也包含很多部分，脓毒症早期预测中最常使用的为护理记录和临床相关记录。临床环境下，急诊常作为脓毒症早期识别的第一站，利用第一手临床信息对患者进行识别就显得尤为重要。Horng 等人[15]将急诊就诊时的主诉、病史特点及护理评估记录等作为非结构化数据，并与容易获取的生命体征等结构化数据相结合，利用支持向量机(SVM)进行模型建立，使模型的预测准确性明显提高(AUC 从 0.67 显著提高到 0.86)，提示非结构化数据可以帮助我们提高预测效能，并有可能在未来开发出相关程序用于急诊快速识别脓毒症；Apostolova 等人[22]开发了一种自动监测护理笔记以识别脓毒症的早期迹象模型，使得 F1 分数(用于衡量分类模型的精确率和召回率)从 79%提升到了 96%。Qin 等人[23]通过结合护理记录和病程记录等文本特征作为预测因子，最后也得出结论，加入文本后可显著提高脓毒症预测模型的性能。

3.2. 儿童脓毒症早期预测模型

儿童领域利用机器学习进行脓毒症预测模型的研究也在逐渐发展中，在诊断预测方面，2022 年 Solé-Ribalta 等人[24]收集西班牙某医院 210 名患儿，45 个预测指标，利用机器学习方法，开发出了 PESERS (Pediatric Sepsis Recognition and Stratification)评分模型，该评分模型精简至 11 个变量，与 PRISM III、pSOFA 和 PELOD-2 等儿童器官功能障碍评分相比，展现出卓越的辨别力和准确性，对败血症的早期检测及严重程度预测具有显著临床意义。同年，Chen 等人[25]利用上海某儿童医院 PICU 数据进行儿童脓毒症早期预测模型建立，利用初次入院检查记录(包含病史记录、实验室检查等)，根据入院后的不同时间段进行风险预测，实现了脓毒症的实时诊断预测，最后得到的新算法可缩短首次抗生素的使用时间，以改善预后。

在脓毒症死亡预测方面，2023 年李少军等人[26]利用单中心 PICU 感染患儿数据对脓毒症及脓毒性休克的死亡风险进行预测。研究收集了患儿数据，经过特征筛选后，运用了多种机器学习方法，包括 LG、LDA、KNN、CART、NB、SVM、RF 和 GBM，建立了预测模型。经过验证，GBM 预测模型在脓毒症死亡预测方面展现出了较高的准确性和区分度，因此被综合评估为最佳模型。

4. 局限性

目前大多数研究是在单中心进行，缺乏普适性，多为针对 ICU 环境下病人，对于 ICU 环境外的脓毒症诊断尚不足。并且研究大多聚焦于中高收入国家，然而，中低收入国家脓毒症的负担实际上更为沉重。此外，大多数研究集中在成人脓毒症，而针对儿童的研究较少，尽管儿童脓毒症的发病率和死亡率都较高。因此，未来的研究需要更多关注中低收入国家儿童脓毒症的诊断标准，以实现早期诊断和积极治疗，改善患儿的预后。最后，考虑到非结构化数据具有数据量大、差异性显著、可解释性低等特点，未来的研究应致力于寻找更先进的数据融合处理方法，如通过深入研究和应用深度学习、自然语言处理、联邦学习、图神经网络以及迁移学习等先进的数据融合处理方法，我们能够更加具体和具备操作性地推动脓毒症预测模型的进展。在未来的研究中，应重点关注这些方法在医疗领域的定制化应用与优化策略，以促进数据融合技术的进一步发展，从而为脓毒症的早期诊断和积极治疗提供有力的支持。

5. 总结

总之，和结构化数据相比，非结构化数据包含更为丰富的临床信息及个性化特征，使得开发出的模型在关键性能指标如准确性及召回率上均有所提升，并且能够提前脓毒症的诊断时间，为治疗提供更佳的指导。因此，利用非结构化数据预测脓毒症的前景是相当乐观的。随着科技的发展，我们有充分的理由相信，未来将有更多转换和融合技术被应用于处理非结构化数据，以开发出更为精确的预测模型，从而提升儿童脓毒症的早期诊断能力，并更好地改善患儿的预后。

参考文献

[1] Rudd, K.E., Johnson, S.C., Agesa, K.M., Shackelford, K.A., Tsoi, D., Kievlan, D.R., *et al*. (2020) Global, Regional, and National Sepsis Incidence and Mortality, 1990-2017: Analysis for the Global Burden of Disease Study. *The Lancet*, **395**, 200-211. https://doi.org/10.1016/s0140-6736(19)32989-7

[2] Kissoon, N. and Uyeki, T.M. (2016) Sepsis and the Global Burden of Disease in Children. *JAMA Pediatrics*, **170**, 107-108. https://doi.org/10.1001/jamapediatrics.2015.3241

[3] Bassat, Q., Blau, D.M., Ogbuanu, I.U., Samura, S., Kaluma, E., Bassey, I., *et al*. (2023) Causes of Death among Infants and Children in the Child Health and Mortality Prevention Surveillance (CHAMPS) Network. *JAMA Network Open*, **6**, e2322494. https://doi.org/10.1001/jamanetworkopen.2023.22494

[4] Marik, P.E. and Farkas, J.D. (2018) The Changing Paradigm of Sepsis: Early Diagnosis, Early Antibiotics, Early Pressors, and Early Adjuvant Treatment. *Critical Care Medicine*, **46**, 1690-1692. https://doi.org/10.1097/ccm.0000000000003310

[5] 王仲, 魏捷, 朱华栋, 等. 中国脓毒症早期预防与阻断急诊专家共识[J]. 中国急救医学, 2020, 40(7): 577-588.

[6] Kumar, A., Roberts, D., Wood, K.E., Light, B., Parrillo, J.E., Sharma, S., *et al*. (2006) Duration of Hypotension before Initiation of Effective Antimicrobial Therapy Is the Critical Determinant of Survival in Human Septic Shock. *Critical Care Medicine*, **34**, 1589-1596. https://doi.org/10.1097/01.ccm.0000217961.75225.e9

[7] Beam, A.L. and Kohane, I.S. (2018) Big Data and Machine Learning in Health Care. *Journal of the American Medical Association*, **319**, 1317-1318. https://doi.org/10.1001/jama.2017.18391

[8] Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., *et al*. (1996) The SOFA (Sepsis-Related Organ Failure Assessment) Score to Describe Organ Dysfunction/Failure. *Intensive Care Medicine*, **22**, 707-710. https://doi.org/10.1007/bf01709751

[9] Levy, M.M., Fink, M.P., Marshall, J.C., Abraham, E., Angus, D., *et al*. (2003) 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. *Intensive Care Medicine*, **29**, 530-538. https://doi.org/10.1007/s00134-003-1662-x

[10] Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., *et al*. (2021) Challenges and Opportunities beyond Structured Data in Analysis of Electronic Health Records. *WIREs Computational Statistics*, **13**, e1549. https://doi.org/10.1002/wics.1549

[11] Zhang, D., Yin, C., Zeng, J., Yuan, X. and Zhang, P. (2020) Combining Structured and Unstructured Data for Predictive Models: A Deep Learning Approach. *BMC Medical Informatics and Decision Making*, **20**, 1-11. https://doi.org/10.1186/s12911-020-01297-6

[12] 吴宗友, 白昆龙, 杨林蕊, 等. 电子病历文本挖掘研究综述[J]. 计算机研究与发展, 2021, 58(3): 513-527.

[13] Spasic, I. and Nenadic, G. (2020) Clinical Text Data in Machine Learning: Systematic Review. *JMIR Medical Informatics*, **8**, e17984. https://doi.org/10.2196/17984

[14] Goh, K.H., Wang, L., Yeow, A.Y.K., Poh, H., Li, K., Yeow, J.J.L., *et al*. (2021) Artificial Intelligence in Sepsis Early Prediction and Diagnosis Using Unstructured Data in Healthcare. *Nature Communications*, **12**, Article No. 711. https://doi.org/10.1038/s41467-021-20910-4

[15] Horng, S., Sontag, D.A., Halpern, Y., Jernite, Y., Shapiro, N.I. and Nathanson, L.A. (2017) Creating an Automated Trigger for Sepsis Clinical Decision Support at Emergency Department Triage Using Machine Learning. *PLOS ONE*, **12**, e0174708. https://doi.org/10.1371/journal.pone.0174708

[16] Alsentzer, E., Murphy, J., Boag, W., Weng, W., Jindi, D., Naumann, T., *et al*. (2019) Publicly Available Clinical. *Proceedings of the* 2*nd Clinical Natural Language Processing Workshop*, Minneapolis, June 2019, 72-78. https://doi.org/10.18653/v1/w19-1909

[17] Amrollahi, F.P., Shashikumar, S., Razmi, F. and Nemati, S. (2021) Contextual Embeddings from Clinical Notes Improves Prediction of Sepsis. *Intensive Care and Critical Care Medicine*, 1-6. https://doi.org/10.1101/2021.03.02.21252779

[18] Barton, C., Chettipally, U., Zhou, Y., Jiang, Z., Lynn-Palevsky, A., Le, S., *et al*. (2019) Evaluation of a Machine Learning Algorithm for up to 48-Hour Advance Prediction of Sepsis Using Six Vital Signs. *Computers in Biology and Medicine*, **109**, 79-84. https://doi.org/10.1016/j.compbiomed.2019.04.027

[19] Mao, Q., Jay, M., Hoffman, J.L., Calvert, J., Barton, C., Shimabukuro, D., *et al*. (2018) Multicentre Validation of a Sepsis Prediction Algorithm Using Only Vital Sign Data in the Emergency Department, General Ward and ICU. *BMJ Open*, **8**, e017833. https://doi.org/10.1136/bmjopen-2017-017833

[20] van Wyk, F., Khojandi, A. and Kamaleswaran, R. (2019) Improving Prediction Performance Using Hierarchical Analysis of Real-Time Data: A Sepsis Case Study. *IEEE Journal of Biomedical and Health Informatics*, **23**, 978-986. https://doi.org/10.1109/jbhi.2019.2894570

[21] Wang, D., Li, J., Sun, Y., Ding, X., Zhang, X., Liu, S., *et al*. (2021) A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients. *Frontiers in Public Health*, **9**, Article 754348. https://doi.org/10.3389/fpubh.2021.754348

[22] Apostolova, E. and Velez, T. (2017) Toward Automated Early Sepsis Alerting: Identifying Infection Patients from Nursing Notes. *BioNLP* 2017, Vancouver, August 2017, 257-262. https://doi.org/10.18653/v1/w17-2332

[23] Qin, F., Madan, V., Ratan, U., *et al*. (2021) Improving Early Sepsis Prediction with Multi Modal Learning.

[24] Solé-Ribalta, A., Launes, C., Felipe-Villalobos, A., Balaguer, M., Luaces, C., Garrido, R., *et al*. (2022) New Multivariable Prediction Model Pediatric Sepsis Recognition and Stratification (PESERS Score) Shows Excellent Discriminatory Capacity. *Acta Paediatrica*, **111**, 1209-1219. https://doi.org/10.1111/apa.16321

[25] Chen, X., Zhang, R. and Tang, X.Y. (2021) Towards Real-Time Diagnosis for Pediatric Sepsis Using Graph Neural Network and Ensemble Methods. *European Review for Medical & Pharmacological Sciences*, **25**, 4693-4701.

[26] 李少军. 基于大数据和机器学习开发儿童脓毒症诊断与预后模型的队列研究[D]: [博士学位论文]. 重庆: 重庆医科大学, 2023.