·述评·

重视大数据分析在脓毒症研究中的应用

刘辉[1] 姚咏明[2]

[1] 中国人民解放军总医院第一医学中心重症医学科，北京 100853；[2] 中国人民解放军总医院医学创新研究部转化医学研究中心，北京 100853

通信作者：姚咏明，Email: c_ff@sina.com

基金项目：国家重点研发计划项目（2022YFA1104600），北京市自然科学基金项目（7222162），国家自然科学基金重点项目（82130062, 82241062）

DOI:10.3760/cma.j.issn.1671-0282.2024.02.001

随着计算机技术以及信息手段的进步，研究者能够对海量的医疗、生物数据进行收集、存储和分析。近年来，大数据分析在脓毒症研究中得到迅速发展和广泛应用。2015 年，全球专家基于 300 余万份电子病例大数据分析，推动了脓毒症定义 3.0（sepsis 3.0）的形成。其中快速序贯器官功能衰竭评分（quick sequential organ failure assessment, qSOFA）利用血压、呼吸频率及意识变化来快速判断患者脓毒症风险，也是基于大数据分析筛选。2017 年，1 项发表在《美国医学会杂志》上的临床观察验证了 SOFA 评分在预后判断上的准确性，该研究分析了来自澳大利亚、新西兰 182 家重症监护病房（intensive care unit, ICU）、184 875 份病例资料[1]。目前，大数据分析在脓毒症病理过程心电、脑电等时序数据、影像学数据、组学数据以及生物信息等海量数据处理中受到极大关注与高度重视。

1 重症电子病历大数据库

来源于医院信息系统的大数据经过去隐私、结构化处理，形成了具有庞大体量的电子病历数据库。目前，较为知名的单中心医疗大数据是美国麻省理工学院参与开发的重症医学信息数据库（medical information mart for intensive care, MIMIC），该数据来源于贝斯以色列女执事医学中心 2001 年至 2012 年 6 万余例 ICU 危重患者的住院信息。另一个著名的重症数据库是重症电子病历合作研究数据库（eICU collaborative research database, eICU-CRD）。eICU-CRD 是飞利浦集团与麻省理工学院计算生理学实验室合作创建的大型公共数据库，属于多中心数据库。eICU-CRD 数据库涵盖了 2014 年和 2015 年入住 ICU 的 20 余万例患者医疗数据。此外，澳大利亚和新西兰的重症监护病房数据库（Australian and New Zealand Intensive Care Adult Patient Data-Base, ANZICS-APD）发布于 2010 年，是最大的 2 个国家医学数据库，超过 200 家 ICU 的数据链接入该数据库，ICU 入院记录超过 200 万条。促进 sepsis 定义更新为 Sepsis 3.0 的“1/8 SIRS 阴性 sepsis”资料正是出自该大型数据库，研究成果发表于《新英格兰医学杂志》[2]。其他免费公开的欧美数据库包括 ICU 高时间分辨率数据库（high time resolution ICU dataset, HiRID）[3] 及阿姆斯特丹大学医学中心数据库（Amsterdam University Medical Centers Database, AmsterdamUMCdb）等[4]。

我国重症医学数据库建设也在快速发展，并于 2023 年发布了重症数据库应用中国专家共识[5]。目前，儿童重症数据库（paediatric intensive care database, PIC）和 ICU 感染患者数据库（critical care database comprising patients with infection）[6] 为国内公开的重症医学数据库，在 PhysioNet 网站上提供检索查询。PIC 数据库收录了浙江大学医学院附属儿童医院在 2010 年至 2018 年期间、共计 12 881 例不同儿科患者 13 499 次住院资料。ICU 感染患者数据库是四川省自贡市第四人民医院的单中心数据库，收录了 2019 年至 2020 年间共 2 790 例 ICU 感染患者的临床资料。另外，其他单中心数据库正在逐步建立，包括北京协和医院、复旦大学附属中山医院、解放军总医院等都在开展相关工作。

2 脓毒症研究中生物信息学大数据

科技的发展使得基因测序成本迅速下降，大

规模基因分析成为现实。当今，高通量基因检测数据量达到了惊人的程度，某些大样本分析可能会动用十几台服务器进行长时间计算。脓毒症研究常用到的生物信息数据库包括基因表达（gene expression omnibus, GEO）数据库，GEO 数据库是由美国国立生物技术信息中心（National Center for Biotechnology Information, NCBI）创建并维护的基因表达数据库 [7]。它创建于 2000 年，收录了由世界各地实验室提交的超过 1 871 121 个样本试验资料，16 088 个芯片平台记录，71 339 种实验项目以及 3 848 种研究类型的基因表达谱数据。还有 GO 数据库，全称是基因本体（gene ontology ），可以提供与目标基因在细胞组分、分子功能和生物过程三个层面上密切相关的基因信息。此外，基因组路径数据库（kyoto encylopaedia of genes and genomes, KEGG）包括大多数已知的代谢途径和部分已知的调控途径；而通过蛋白质相互作用网络（protein-protein interaction network, PPI）有助于挖掘核心的调控基因，其中最著名的 PPI 网络 string 网站（https://string-db.org/）存储了 2 031 个物种、9 643 763 种蛋白分子，共 1 380 838 440 个相互作用信息。其他的生物信息学数据库还有很多，可以对目标基因、蛋白分子、免疫微环境、单细胞浸润等不同方面进行深入探讨，为脓毒症研究提供了重要数据。例如，有研究比较了脓毒症患者与正常对照人群的基因表达差异，筛选出 101 个差异表达基因（differentially expressed genes, DEGs），将筛选出来的 DEGs 进一步分析，通过 GO 数据库发现这些上调的基因主要与干扰素及粒细胞分化调控有关；通过 KEGG 数据库揭示其相关信号通路集中在补体调节、凝血机制激活方面 [8]。另有学者探索了新冠感染、急性呼吸窘迫综合征以及脓毒症患者发病的内在共同机制。基因差异分析显示 3 种疾病状态下有 110 个共同表达基因，其中 4 个基因为中心基因（hub gene），并且通过蛋白 - 药物相互作用（protein-drug interaction, PDI）数据库筛选出异氟泼尼松等 10 种可能有效治疗药物 [9]。总之，脓毒症可能由多种疾病引起，具有高度异质性，高通量基因测序技术以及生物信息学分析为深刻解析其发病本质开辟了新途径。

3 组学大数据分析

组学大数据是指基于高通量检测方法得到的体量庞大的某一类数据信息。其中，基因组学关注基因结构与功能分析，基因多态性可造成脓毒症患者在炎症因子、细胞受体以及信号通路方面的差异。而表观基因组学则针对基因修饰与基因表达之间的关系，基因修饰改变并不影响脱氧核苷酸（deoxyribonucleic acid, DNA）的序列结构，但是对基因表达产生较大影响。而转录组学针对的是核糖核酸（ribonucleic acid, RNA）和非编码 RNA，关注基因表达的活性与调节。通过基因相关组学的研究，可以从微观水平甄别不同类型脓毒症发病机制。例如，免疫麻痹是脓毒症患者病情恶化的关键因素，系脓毒症领域的前沿热点问题。通过单细胞转录组分析，我们发现并鉴定了一群具有免疫抑制功能 HLA-DRlowS100Ahigh 单核细胞，为脓毒症个体化免疫调理提供了新方案 [10]。另据报道，脓毒症时循环中单核细胞基因的甲基化组以及组蛋白乙酰化发生广泛变化，并导致细胞因子 IL-10、IL-6 表达异常，其改变与器官功能损伤相关 [11-12]。此外，在微生物群落分析中使用高通量测序催生了一个新的科学领域，就是宏基因组。宏基因组学是指应用测序技术来分析样本中存在的全部基因组信息 [13]。蛋白组学与脓毒症患者预警、诊断和预后等生物标志物关系紧密。通过质谱分析发现，波形蛋白（viment）水平在脓毒症及脓毒性休克患者中明显升高，其中死亡组血清含量更高，而波形蛋白可参与调控淋巴细胞凋亡和炎症反应，可能成为脓毒症预后的生物标志物 [14]。代谢组学（metabonomics/metabolomics）是 20 世纪 90 年代末期发展起来的一门新兴学科，是研究有关生物体被扰动后（如基因改变或环境变化）其代谢产物（内源性代谢物质）种类、数量及其变化规律的科学。不同类型脓毒症患者的代谢组学分析表型也不一样 [15]，通过核磁共振波谱法、液相色谱及气相色谱 - 质谱法等手段进行大量筛选，能够确定不同下呼吸道感染患儿以及正压通气所致代谢产物之间分布差异，具有较好的诊断评估价值 [16]。另一项研究显示，应用二甲胺、甘露糖等 7 种代谢产物可较好预测患儿脓毒症的严重程度，其预测的曲线下面积（area under curve, AUC）达到 93% [17]。值得指出的是，影像组学新近迅速发展，通过捕获计算机 X 线层成像（computerized tomography, CT）、磁共振成像（magnetic resonance imaging, MRI）、超声等不同来源图像的组织和病变特征，如异质性、形状等，形

成全方位、多维度信息，将对脓毒症潜在特点、变化规律以及治疗模式产生深远影响。

4 多模态大数据分析

大数据来源是多样的、多维度的，其整合分析手段突破了传统研究方法的局限性。多模态数据丰富了研究对象的多维度、互补信息，有力促进其研究论证过程[18-19]。其来源除了传统的临床试验、基础医学实验，还包括医院电子病历、医学影像或者可穿戴设备、以及其他健康相关行为产生的大数据；还可分为诸如类别、重量、身高等静态数据，以及包括不同时间点采集的血糖、体温，甚至不同时长的超声、心电等动态数据。

2019 年，斯坦福大学与麻省理工学院发布了两个大型医疗公开数据集：CheXpert 和 MIMIC-CXR，其中 CheXpert 内含 224 316 张 X 光胸片、MIMIC-CXR 内含 371 920 张带标签的胸片。两个数据集的数据量级和标注精准度都非常高[20-21]。有研究利用该数据库，通过深度学习的大数据挖掘方法对患者胸片、影像学资料以及临床结构化数据进行综合分析，可以对脓毒症患者的微生物学类型进行预测。对真菌、细菌及病毒的预测效能（即 AUC））分别达到 0.81、0.83 和 0.79[22]。还有学者分析了 MIMIC-CXR 中 64 581 例患者 369 071 张胸片和相关诊断报告，对胸片肺水肿的严重程度进行了准确分级[23]。值得指出的是，对于动态数据的分析，大数据分析相关方法亦具有良好时序性捕获能力，较为经典的是长短期记忆网络（long short term memory, LSTM）和门控循环单元（gated recurrent unit, GRU）等循环神经网络算法。可对各动态指标单独建模，捕获其数值变化、检测频率变化，去除指标数据不规则性和稀疏性对预测模型的影响，并能够联合静态指标的特征，再分别挖掘不同模态的关键信息，最终将重构后的信息融合，获得患者健康风险评估结果[24]。一项研究针对大量心电图数据进行分析及模型训练，包括 44 959 例心功能不全患者的心电图资料，得到了预测左心功能障碍的算法模型，其预测效能的敏感性、特异性、精确度分别达到 86.3%、85.7% 和 85.7%[25]。

5 研究展望

脓毒症研究中大数据分析的应用具有知识驱动和数据密集等特点，其快速发展是建立在临床医师、数据工程师以及计算机科学家之间良好协作基础之上。临床专家提出科学问题，而数据和计算机工程师提供解决方案。当前，既熟悉脓毒症研究进展、又掌握大数据分析方法的交叉融合型人才成为推动脓毒症大数据应用的关键。其次，许多脓毒症领域的大数据分析方法和技术可相互借鉴。例如，生物信息学最早聚焦于肿瘤研究领域，差异基因表达、功能分析等成熟方法近年来在脓毒症研究中得以应用，拓展了脓毒症领域的深度和广度。时至今日，多种大数据分析和人工智能技术也在不断融合发展，例如基因影像组学，将微观与宏观的信息进行结合；还有多组学整合分析技术，研究者可从基因组、转录组、代谢组等不同层面获取数据信息。

总之，大数据分析的发展得益于计算机科学进步，特别在近十年算力快速提升后得到迅猛发展和广泛应用。相信大数据分析策略革新将有力推动与促进脓毒症研究不断向前发展。

利益冲突 所有作者声明无利益冲突

参 考 文 献

[1] Raith EP, Udy AA, Bailey M, et al. Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for In-hospital mortality among adults with suspected infection admitted to the intensive care unit[J]. JAMA, 2017, 317(3): 290-300. DOI: 10.1001/jama.2016.20328.

[2] Kaukonen KM, Bailey M, Pilcher D, et al. Systemic inflammatory response syndrome criteria in defining severe sepsis[J]. N Engl J Med, 2015, 372(17): 1629-1638. DOI: 10.1056/nejmoa1415236.

[3] Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning[J]. Nat Med, 2020, 26(3): 364-373. DOI: 10.1038/s41591-020-0789-4.

[4] Thoral PJ, Peppink JM, Driessen RH, et al. Sharing ICU patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: the Amsterdam university medical centers database (AmsterdamUMCdb) example[J]. Crit Care Med, 2021, 49(6): e563-e577. DOI: 10.1097/CCM.0000000000004916.

[5] 中国卫生信息与健康医疗大数据学会重症医学分会, 北京肿瘤学会重症医学专业委员会. 重症大数据应用中国专家共识(2022)[J]. 中华医学杂志, 2023, 103(6): 404-424. DOI: 10.3760/cma.j.cn112137-20221008-02098.

[6] Xu P, Chen L, Zhu YF, et al. Critical care database comprising patients

with infection[J]. Front Public Health, 2022, 10: 852410. DOI: 10.3389/fpubh.2022.852410.

[7] Clough E, Barrett T. The gene expression omnibus database[J]. Methods Mol Biol, 2016, 1418: 93-110. DOI: 10.1007/978-1-4939-3578-9_5.

[8] Li YR, Zhang HY, Shao JY, et al. Bioinformatics analysis for identifying pertinent pathways and genes in sepsis[J]. Comput Math Methods Med, 2021, 2021: 2085173. DOI: 10.1155/2021/2085173.

[9] Li PY, Li T, Zhang ZM, et al. Bioinformatics and system biology approach to identify the influences among COVID-19, ARDS and sepsis[J]. Front Immunol, 2023, 14: 1152186. DOI: 10.3389/fimmu.2023.1152186.

[10] Yao RQ, Zhao PY, Li ZX, et al. Single-cell transcriptome profiling of sepsis identifies HLA-DRlowS100A high monocytes with immunosuppressive function[J]. Mil Med Res, 2023, 10(1): 27. DOI: 10.1186/s40779-023-00462-y.

[11] Lorente-Sorolla C, Garcia-Gomez A, Català-Moll F, et al. Inflammatory cytokines and organ dysfunction associate with the aberrant DNA methylome of monocytes in sepsis[J]. Genome Med, 2019, 11(1): 66. DOI: 10.1186/s13073-019-0674-2.

[12] Zheng ZH, Huang G, Gao T, et al. Epigenetic changes associated with interleukin-10[J]. Front Immunol, 2020, 11: 1105. DOI: 10.3389/fimmu.2020.01105.

[13] Lema NK, Gemeda MT, Woldesemayat AA. Recent advances in metagenomic approaches, applications, and challenge[J]. Curr Microbiol, 2023, 80(11): 347. DOI: 10.1007/s00284-023-03451-5.

[14] Su LX, Pan P, Yan P, et al. Role of vimentin in modulating immune cell apoptosis and inflammatory responses in sepsis[J]. Sci Rep, 2019, 9(1): 5747. DOI: 10.1038/s41598-019-42287-7.

[15] Hussain H, Vutipongsatorn K, Jiménez B, et al. Patient stratification in sepsis: using metabolomics to detect clinical phenotypes, sub-phenotypes and therapeutic response[J]. Metabolites, 2022, 12(5): 376. DOI: 10.3390/metabo12050376.

[16] Wildman E, Mickiewicz B, Vogel HJ, et al. Metabolomics in pediatric lower respiratory tract infections and sepsis: a literature review[J]. Pediatr Res, 2023, 93(3): 492-502. DOI: 10.1038/s41390-022-02162-0.

[17] Mickiewicz B, Thompson GC, Blackwood J, et al. Biomarker phenotype for early diagnosis and triage of sepsis to the pediatric intensive care unit[J]. Sci Rep, 2018, 8(1): 16606. DOI: 10.1038/s41598-018-35000-7.

[18] Bayoudh K, Knani R, Hamdaoui F, et al. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets[J]. Vis Comput, 2022, 38(8): 2939-2970. DOI: 10.1007/s00371-021-02166-7.

[19] Jeon G, Bellandi V, Chehri A, et al. Big multimodal data analysis: models and performance analysis[J]. Big Data, 2022, 10(5): 369-370. DOI: 10.1089/big.2022.0216.

[20] Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study[J]. Lancet Digit Health, 2022, 4(6): e406-e414. DOI: 10.1016/S2589-7500(22)00063-2.

[21] Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports[J]. Sci Data, 2019, 6(1): 317. DOI: 10.1038/s41597-019-0322-0.

[22] Boussina A, Ramesh K, Arora H, et al. Differentiation of fungal, viral, and bacterial sepsis using multimodal deep learning[J]. medRxiv, 2023: 2023.04.10.23288378. DOI: 10.1101/2023.04.10.23288378.

[23] Horng S, Liao RZ, Wang X, et al. Deep learning to quantify pulmonary edema in chest radiographs[J]. Radiol, 2021, 3(2): e190228. DOI: 10.1148/ryai.2021190228.

[24] Rafiei A, Rezaee A, Hajati F, et al. SSP: early prediction of sepsis using fully connected LSTM-CNN model[J]. Comput Biol Med, 2021, 128: 104110. DOI: 10.1016/j.compbiomed.2020.104110.

[25] Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram[J]. Nat Med, 2019, 25(1): 70-74. DOI: 10.1038/s41591-018-0240-2.

（收稿日期：2023-12-21）

（本文编辑：何小军）