·专论·

医疗大数据在神经系统疾病中的应用

冯铭　郑雪晴　王任直

【摘要】 神经系统疾病病情复杂、种类繁多，患者预后相当程度取决于医师的临床经验和诊疗水平。近年随着医疗大数据的指数式增长与统计分析方法的发展，越来越多的研究通过挖掘并分析大数据以揭示疾病发病机制，辅助临床诊断、决策和治疗，有望提高神经系统疾病的诊疗与预后预测能力。本文从结构化数据、影像组学和生物信息数据三方面阐述医疗大数据在神经系统疾病中的应用。

【关键词】 神经系统疾病；　自动数据处理；　电子健康病历；　计算生物学；　综述

The application of medical big data in central nervous system diseases

FENG Ming, ZHENG Xue-qing, WANG Ren-zhi

Department of Neurosurgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China

Corresponding author: FENG Ming (Email: fengming@pumch.cn)

【Abstract】 Nervous system diseases are associated with complexity and diversity. As a result, the prognosis of patients deeply depends on physicians' clinical experience and their diagnosis and treatment level. With the exponential growth of medical big data and the development of statistical analysis tools, more researchers discover the pathogenesis of diseases and assist clinical process through the mining of big data, so as to improve the diagnosis, treatment, and prognosis of nervous system diseases. This paper elaborates the application of three kinds of medical big data in nervous system diseases, including structured data, radiomics, and bioinformatics.

【Key words】 Nervous system diseases;　Automatic data processing;　Electronic health records;　Computational biology;　Review

This study was supported by Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (No. 2020-I2M-C&T-B-031) and Educational Reform Project of Peking Union Medical College (No. 10023201900107).

Conflicts of interest: none declared

随着信息学和统计学的发展，“大数据”概念兴起并在各领域中展现出其应用价值。大数据的3项核心特征为体量庞大(volumn)、生成速度迅速(velocity)、数据类型丰富(variety)[1]。医疗领域的大数据包括生物信息数据(如基因组学、蛋白质组学、代谢组学等)、影像组学数据(如MRI、CT、分子影像、病理影像等)、结构化数据(如检验结果、诊断、药物治疗等)、非结构化数据(如临床记录)等[1]。采用多种数据挖掘工具对医疗大数据进行开发和分析将成为传统医学模式向精准医学转变的核心动力，医疗大数据的广泛应用也将使人们对健康和疾病的理解产生深远影响。目前，医疗大数据的应用方向主要包括通过机器学习(ML)辅助临床决策、阐释特殊疾病机制、支持药品和医疗机器人等研发、个体化诊疗、重大疾病相关危险因素筛查和风险预测、传染性疾病监测等[2]。神经系统疾病种类繁多，有神经系统肿瘤、脑血管病、脑功能性疾病等，诊断与治疗相对复杂，预后较差[3]。因此，早期诊断与鉴别诊断至关重要，目前迫切需要提高临床决策能力以及精确预防与治疗水平，而基于医疗大数据的分析和应用则提供了新的思路和方法。

一、结构化数据在神经系统疾病中的应用

电子病历(EHR)是由医疗保健者生成并维护

doi:10.3969/j.issn.1672-6731.2021.03.002

基金项目：中国医学科学院医学与健康科技创新工程项目(项目编号：2020-I2M-C&T-B-031)；北京协和医学院教育教学改革项目(项目编号：10023201900107)

作者单位：100730 中国医学科学院 北京协和医学院 北京协和医院神经外科

通讯作者：冯铭，Email：fengming@pumch.cn

的患者健康和临床护理记录，旨在系统收集信息用于更全面精准的临床护理。随着电子病历系统在全世界范围内的日益普及，对其中的高通量真实世界信息进行提取和分析成为可能。电子病历的结构显著影响数据的可用性，结构化数据一致且易于提取，是目前研究的主流；非结构化数据需自然语言处理（NLP）等工具进行标准化、编码和提取，较少用于大数据分析[4]。将机器学习与结构化数据相结合，可以用于垂体腺瘤预后的预测，通过筛选结构化临床特征并开发算法模型，可以用于肢端肥大症早期缓解和库欣病延迟缓解的预测，以指导临床决策[5-6]。但也有部分针对鞍区疾病的机器学习模型选择随意、未提供重复研究所需的参数和超参数、缺乏验证，导致研究结果可重复性、鲁棒性和可泛化性受到限制[7]。脑卒中的结构化数据挖掘已取得一定成果，通过机器学习从电子病历中评估缺血性卒中严重程度的主要评价指标，计算得出美国国立卫生研究院卒中量表（NIHSS）评分是较准确的评价指标[8]；还通过电子病历信息拟合缺血性卒中TOAST分型标准，最终获得预测阳性值达95%的特征提取算法，从而辅助临床上缺血性卒中亚型的准确分类[9]，对于药物治疗、风险评估和二级预防具有重要意义。电子病历的数据挖掘还可用于阿尔茨海默病的发病风险评估、预后预测、临床护理等多方面，发现红细胞沉降率（ESR）与发病风险显著相关[10]；同时还发现首次就诊连线测验-A（TMT-A）评分与疾病进展显著相关[11]，连同其他神经心理学测验的基线特征，有助于预测预后。由此可见，电子病历系统蕴含大量可供学习的数据，但进一步投入临床应用仍需改善不同卫生系统之间电子病历的可获取性、标准化和互用性。电子病历数据不同于研究型数据库，缺乏准确性和完整性，从而限制其研究结果的准确性；此外，对于非结构化数据的整理也将在未来扩展电子病历信息的应用。

二、影像组学在神经系统疾病中的应用

医学影像学作为临床常用的诊断工具，包含大量可供挖掘的信息，其数字化特征也使其具有大数据处理的可能。将生物医学信息中的组学概念迁移至医学影像即形成影像组学，从高通量的医学影像数据中提取深度特征，通过机器学习进行定量分析，而辅助疾病的早期筛查、准确诊断、分级分期、治疗预后和分子特征分析。影像组学将需用于诊断的图像转换为可挖掘的数据，主要包括以下5个步骤，图像采集与重建、兴趣区（ROI）分割与标记、特征提取与量化、统计分析、预测模型建立[3]，这种低成本、非侵入性的动态监测技术对于神经系统疾病优势显著。影像组学特征可以用于脑肿瘤的鉴别诊断[3]，可资鉴别胶质母细胞瘤与中枢神经系统淋巴瘤和脑转移瘤、恶性血管外皮细胞瘤与血管型脑膜瘤。在胶质瘤的诊断与治疗方面，通过机器学习和特征提取并结合影像组学方法，可以精确分级并根据不同级别辅助临床决策；通过对重要分子生物学标志物的分析，如Ki-67抗原标记指数、异柠檬酸脱氢酶（*IDH*）、1p/19q共缺失、端粒酶逆转录酶（*TERT*）、同源性磷酸酶-张力蛋白（*PTEN*）、表皮生长因子受体（*EGFR*）、骨膜蛋白（*POSTN*）、X连锁α地中海贫血伴精神发育迟滞综合征蛋白（*ATRX*）、*TP53*基因突变以及O^6-甲基鸟嘌呤-DNA甲基转移酶（*MGMT*）甲基化等[3]，也可辅助诊断分子亚型。此外，影像组学还隐含疾病的遗传异质性，可揭示肿瘤基因的表达，为基因分型提供无创性的检测手段[12]。基于影像组学的机器学习模型目前还用于术前脑膜瘤分级[13]、侵袭性功能性垂体腺瘤手术效果预测[14]、肢端肥大症患者肿瘤一致性评估[15]和放疗效果预测[16]等。在脑血管病诊断与治疗方面，基于影像组学的机器学习模型可准确鉴别诊断颅内动-静脉畸形与其他病因引起的脑内血肿[17]，亦可用于预测脑出血周围水肿和血肿扩大[18]。对于脑功能性疾病，基于影像组学的机器学习模型可有效识别早期外观正常的脑白质病变[19]、诊断特发性帕金森病和阿尔茨海默病，还可基于定量的生物学标志物，辅助精神分裂症的个体化诊断[20]以及帕金森病的预后预测[21]。由此可见，影像组学可用于不同神经系统疾病的鉴别及分型诊断、分子特征分析、治疗和预后评估，其作为一种低成本的新型临床检测工具可改进神经系统疾病的治疗决策。然而，影像组学广泛应用于临床实践前仍存在挑战：不同来源的影像学数据需经过归一化预处理以提高参数的准确性；精准且快速的图像分割已成为影像组学的瓶颈；机器学习的开发和验证依靠多中心的协作和数据库的建设；对机器学习算法的认识不足使其结果的可解释性受到限制。相信随着机器学习的不断发展，未来影像组学可在神经系统疾病的常规治疗中有更广泛的应用。

三、生物信息学分析在神经系统疾病中的应用

医疗领域的大数据起源于微观组学。随着高

通量杂交阵列技术的快速发展，各种生物信息数据库相继建立，为共享数据提供便捷。生物信息大数据着眼于分子层面，结合临床表象，可加深对疾病发病机制的理解，为精准医学、转化医学带来新的发展机遇[22]。目前已发现垂体腺瘤的诱因和易感基因包括*USP8*、*AIP*、*MEN1*、*CDKN1B*等，其中，*USP8*基因在库欣病中的突变率高达40%～62%，导致去泛素化酶活性增强，抑制EGFR泛素化，使EGFR不断积累诱发肿瘤[23]，不仅揭示了库欣病的分子发病机制，而且提供了一系列治疗靶点。更多针对胶质瘤的数据库，如中国脑胶质瘤基因组学图谱计划（CGGA）、GliomaDB等数据库相继建立，为精准医学的发展奠定数据基础。基于肿瘤基因组学图谱计划（TCGA）分析线粒体丙酮酸载体蛋白1（MPC1）表达变化与预后的关系，*IDH*突变的胶质瘤患者MPC1过表达与更好的总体生存率相关[24]，MPC1表达降低的胶质母细胞瘤患者则总体生存情况较差，并且对替莫唑胺有抗药性的胶质母细胞瘤*MPC1*基因缺失比例较高[25]。针对脑血管病的全基因组关联研究（GWAS）共确定32个与缺血性卒中及其亚型相关的基因位点[26]。联合进行蛋白质组学、代谢组学、转录组学和基因组学等分析，获得缺血性卒中分型、诊断和预后预测的相关生物学标志物[27]，有助于加深对脑卒中病理生理学机制的理解，为疾病的诊断与治疗提供新的思路。然而，在这些生物信息大数据应用于临床实践前，还需经过更多样化的验证，尤其需要扩大非洲地区高质量、全面、准确的表型和基因组学数据[28-29]。代谢组学可用于评估帕金森病不同发展阶段的病理生理学过程，以尽早纠正异常代谢，为个体化药物治疗增加新的可能。多种微观组学数据的挖掘确定至少19个与阿尔茨海默病发病机制相关的蛋白质靶点，且这些靶点均与获批上市或正在进行临床试验的药物相关，证实了组学研究对探究发病机制和药物研发的作用[30]。多种微观组学的结合对精准医学有广阔的发展前景，但也面临一定的挑战，数据混杂因素多、异质性强；数据标准化水平仍需提升；统计学分析技术在人口规模上的应用仍需改进；分析结果难以区分相关性和因果性等。

四、结论

近年来，医疗领域出现可用数据体量、速度和种类的爆炸式增长，越来越多的机器学习应用于医疗大数据的挖掘与分析，在生物学标志物探寻、疾病机制阐明、疗效和预后预测等方面均取得一定的成果，有望成为临床决策的有力辅助工具。神经系统疾病病情复杂、种类繁多，亟待这样一种简单易行的方式提高临床决策能力和精准治疗水平。多模态数据的交叉与融合是大势所趋，目前已有越来越多的研究将影像组学、生物信息数据和电子病历数据相结合进行深度分析。未来尚待进一步建立数据的协作网络、提升数据质量和对数据的分析能力、加强隐私保护与数据安全，充分体现医疗大数据的价值。

利益冲突 无

参 考 文 献

[1] Zhang R, Simon G, Yu F. Advancing Alzheimer's research: a review of big data promises[J]. Int J Med Inform, 2017, 106:48-56.

[2] Gong MC, Lu L. Research progress and application prospect of medical big data[J]. Yi Xue Xin Xi Xue Za Zhi, 2016, 37:9-15.[弓孟春, 陆亮. 医学大数据研究进展及应用前景[J]. 医学信息学杂志, 2016, 37:9-15.]

[3] Fan Y, Feng M, Wang R. Application of radiomics in central nervous system diseases: a systematic literature review[J]. Clin Neurol Neurosurg, 2019, 187:105565.

[4] Abul - Husn NS, Kenny EE. Personalized medicine and the power of electronic health records[J]. Cell, 2019, 177:58-69.

[5] Qiao N, Shen M, He W, He M, Zhang Z, Ye H, Li Y, Shou X, Li S, Jiang C, Wang Y, Zhao Y. Machine learning in predicting early remission in patients after surgical treatment of acromegaly: a multicenter study[J]. Pituitary, 2021, 24:53-61.

[6] Fan Y, Li Y, Bao X, Zhu H, Lu L, Yao Y, Li Y, Su M, Feng F, Feng S, Feng M, Wang R. Development of machine learning models for predicting postoperative delayed remission in patients with Cushing's disease[J]. J Clin Endocrinol Metab, 2021, 106:e217-231.

[7] Qiao N. A systematic review on machine learning in sellar region diseases: quality and reporting items [J]. Endocr Connect, 2019, 8:952-960.

[8] Kogan E, Twyman K, Heap J, Milentijevic D, Lin JH, Alberts M. Assessing stroke severity using electronic health record data: a machine learning approach[J]. BMC Med Inform Decis Mak, 2020, 20:8.

[9] Guan W, Ko D, Khurshid S, Trisini Lipsanopoulos AT, Ashburner JM, Harrington LX, Rost NS, Atlas SJ, Singer DE, McManus DD, Anderson CD, Lubitz SA. Automated electronic phenotyping of cardioembolic stroke[J]. Stroke, 2021, 52:181-189.

[10] Li L, Ruau D, Chen R, Weber S, Butte AJ. Systematic identification of risk factors for Alzheimer's disease through shared genetic architecture and electronic medical records[J]. Pac Symp Biocomput, 2013: 224-235.

[11] Parikh M, Hynan LS, Weiner MF, Lacritz L, Ringe W, Cullum CM. Single neuropsychological test scores associated with rate of cognitive decline in early Alzheimer disease [J]. Clin Neuropsychol, 2014, 28:926-940.

[12] Jain R, Poisson LM, Gutman D, Scarpace L, Hwang SN, Holder CA, Wintermark M, Rao A, Colen RR, Kirby J, Freymann J, Jaffe CC, Mikkelsen T, Flanders A. Outcome prediction in

patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor[J]. Radiology, 2014, 272:484-493.

[13] Park YW, Oh J, You SC, Han K, Ahn SS, Choi YS, Chang JH, Kim SH, Lee SK. Radiomics and machine learning may accurately predict the grade and histological subtype in meningiomas using conventional and diffusion tensor imaging [J]. Eur Radiol, 2019, 29:4068-4076.

[14] Fan Y, Liu Z, Hou B, Li L, Liu X, Liu Z, Wang R, Lin Y, Feng F, Tian J, Feng M. Development and validation of an MRI-based radiomic signature for the preoperative prediction of treatment response in patients with invasive functional pituitary adenoma[J]. Eur J Radiol, 2019, 121:108647.

[15] Fan Y, Hua M, Mou A, Wu M, Liu X, Bao X, Wang R, Feng M. Preoperative noninvasive radiomics approach predicts tumor consistency in patients with acromegaly: development and multicenter prospective validation [J]. Front Endocrinol (Lausanne), 2019, 10:403.

[16] Fan Y, Jiang S, Hua M, Feng S, Feng M, Wang R. Machine learning-based radiomics predicts radiotherapeutic response in patients with acromegaly [J]. Front Endocrinol (Lausanne), 2019, 10:588.

[17] Zhang Y, Zhang B, Liang F, Liang S, Zhang Y, Yan P, Ma C, Liu A, Guo F, Jiang C. Radiomics features on non-contrast-enhanced CT scan can precisely classify AVM-related hematomas from other spontaneous intraparenchymal hematoma types[J]. Eur Radiol, 2019, 29:2157-2165.

[18] Xie H, Ma S, Wang X, Zhang X. Noncontrast computer tomography-based radiomics model for predicting intracerebral hemorrhage expansion: preliminary findings and comparison with conventional radiological model[J]. Eur Radiol, 2020, 30: 87-98.

[19] Shao Y, Chen Z, Ming S, Ye Q, Shu Z, Gong C, Pang P, Gong X. Predicting the development of normal-appearing white matter with radiomics in the aging brain: a longitudinal clinical study [J]. Front Aging Neurosci, 2018, 10:393.

[20] Cui LB, Liu L, Wang HN, Wang LX, Guo F, Xi YB, Liu TT, Li C, Tian P, Liu K, Wu WJ, Chen YH, Qin W, Yin H. Disease definition for schizophrenia by functional connectivity using radiomics strategy[J]. Schizophr Bull, 2018, 44:1053-1059.

[21] Rahmim A, Huang P, Shenkov N, Fotouhi S, Davoodi-Bojd E, Lu L, Mari Z, Soltanian-Zadeh H, Sossi V. Improved prediction of outcome in Parkinson's disease using radiomics analysis of longitudinal DAT SPECT images[J]. Neuroimage Clin, 2017, 16:539-544.

[22] Wu ZB. Promote translational medicine research of pituitary adenoma in the era of big data[J]. Zhonghua Yi Xue Za Zhi, 2016, 96:1473-1474.[吴哲褒. 促进大数据时代垂体腺瘤转化医学研究[J]. 中华医学杂志, 2016, 96:1473-1474.]

[23] Reincke M, Sbiera S, Hayakawa A, Theodoropoulou M, Osswald A, Beuschlein F, Meitinger T, Mizuno-Yamasaki E, Kawaguchi K, Saeki Y, Tanaka K, Wieland T, Graf E, Saeger W, Ronchi CL, Allolio B, Buchfelder M, Strom TM, Fassnacht M, Komada M. Mutations in the deubiquitinase gene USP8 cause Cushing's disease[J]. Nat Genet, 2015, 47:31-38.

[24] Karsy M, Guan J, Huang LE. Prognostic role of mitochondrial pyruvate carrier in isocitrate dehydrogenase-mutant glioma[J]. J Neurosurg, 2018, 130:56-66.

[25] Chai Y, Wang C, Liu W, Fan Y, Zhang Y. MPC1 deletion is associated with poor prognosis and temozolomide resistance in glioblastoma[J]. J Neurooncol, 2019, 144:293-301.

[26] Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, Rutten-Jacobs L, Giese AK, van der Laan SW, Gretarsdottir S, Anderson CD, Chong M, Adams HHH, Ago T, Almgren P, Amouyel P, Ay H, Bartz TM, Benavente OR, Bevan S, Boncoraglio GB, Brown RD Jr, Butterworth AS, Carrera C, Carty CL, Chasman DI, Chen WM, Cole JW, Correa A, Cotlarciuc I, Cruchaga C, Danesh J, de Bakker PIW, DeStefano AL, den Hoed M, Duan Q, Engelter ST, Falcone GJ, Gottesman RF, Grewal RP, Gudnason V, Gustafsson S, Haessler J, Harris TB, Hassan A, Havulinna AS, Heckbert SR, Holliday EG, Howard G, Hsu FC, Hyacinth HI, Ikram MA, Ingelsson E, Irvin MR, Jian X, Jiménez-Conde J, Johnson JA, Jukema JW, Kanai M, Keene KL, Kissela BM, Kleindorfer DO, Kooperberg C, Kubo M, Lange LA, Langefeld CD, Langenberg C, Launer LJ, Lee JM, Lemmens R, Leys D, Lewis CM, Lin WY, Lindgren AG, Lorentzen E, Magnusson PK, Maguire J, Manichaikul A, McArdle PF, Meschia JF, Mitchell BD, Mosley TH, Nalls MA, Ninomiya T, O'Donnell MJ, Psaty BM, Pulit SL, Rannikmäe K, Reiner AP, Rexrode KM, Rice K, Rich SS, Ridker PM, Rost NS, Rothwell PM, Rotter JI, Rundek T, Sacco RL, Sakaue S, Sale MM, Salomaa V, Sapkota BR, Schmidt R, Schmidt CO, Schminke U, Sharma P, Slowik A, Sudlow CLM, Tanislav C, Tatlisumak T, Taylor KD, Thijs VNS, Thorleifsson G, Thorsteinsdottir U, Tiedt S, Trompet S, Tzourio C, van Duijn CM, Walters M, Wareham NJ, Wassertheil-Smoller S, Wilson JG, Wiggins KL, Yang Q, Yusuf S, Bis JC, Pastinen T, Ruusalepp A, Schadt EE, Koplev S, Björkegren JLM, Codoni V, Civelek M, Smith NL, Trégouët DA, Christophersen IE, Roselli C, Lubitz SA, Ellinor PT, Tai ES, Kooner JS, Kato N, He J, van der Harst P, Elliott P, Chambers JC, Takeuchi F, Johnson AD, Sanghera DK, Melander O, Jern C, Strbian D, Fernandez-Cadenas I, Longstreth WT Jr, Rolfs A, Hata J, Woo D, Rosand J, Pare G, Hopewell JC, Saleheen D, Stefansson K, Worrall BB, Kittner SJ, Seshadri S, Fornage M, Markus HS, Howson JMM, Kamatani Y, Debette S, Dichgans M; AFGen Consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, International Genomics of Blood Pressure (iGEN-BP) Consortium, INVENT Consortium, STARNET, BioBank Japan Cooperative Hospital Group, COMPASS Consortium, EPIC-CVD Consortium, EPIC-InterAct Consortium, International Stroke Genetics Consortium (ISGC), METASTROKE Consortium, Neurology Working Group of the CHARGE Consortium, NINDS Stroke Genetics Network (SiGN), UK Young Lacunar DNA Study, MEGASTROKE Consortium. Multiancestry genome-wide association study of 520, 000 subjects identifies 32 loci associated with stroke and stroke subtypes[J]. Nat Genet, 2018, 50:524-537.

[27] Montaner J, Ramiro L, Simats A, Tiedt S, Makris K, Jickling GC, Debette S, Sanchez JC, Bustamante A. Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke [J]. Nat Rev Neurol, 2020, 16:247-264.

[28] Owolabi M, Peprah E, Xu H, Akinyemi R, Tiwari HK, Irvin MR, Wahab KW, Arnett DK, Ovbiagele B. Advancing stroke genomic research in the age of Trans-Omics big data science: emerging priorities and opportunities[J]. J Neurol Sci, 2017, 382:18-28.

[29] Wahab KW, Tiwari HK, Ovbiagele B, Sarfo F, Akinyemi R, Traylor M, Rotimi C, Markus HS, Owolabi M. Genetic risk of spontaneous intracerebral hemorrhage: systematic review and future directions[J]. J Neurol Sci, 2019, 407:116526.

[30] Sancesario GM, Bernardini S. Alzheimer's disease in the omics era[J]. Clin Biochem, 2018, 59:9-16.

（收稿日期：2021-03-16）
（本文编辑：彭一帆）