［专题组稿专家简介］ 薛万国，解放军总医院医学信息情报所高级工程师，中国医院协会信息管理专业委员会副主任委员、中国卫生信息学会常务理事、北京市信息化专家咨询委员会委员。从事医院信息化研究开发与推广应用工作 30 年，作为主要人员承担了“军字一号”医院信息系统的研发工作，在国内较早系统化地开展电子病历研究，牵头多项卫生信息标准和规范的制订，曾获得国家科技进步二等奖等奖项。现负责解放军总医院医疗大数据中心的建设工作。

大数据时代的医学创新与现实挑战

薛万国，应 俊
解放军总医院 医疗大数据中心，北京 100853

摘要：大数据背景下，医学研究所依赖的数据环境和分析技术发生了很大变化。以机器学习、深度学习等技术为代表的预测型分析和指导型分析突破了传统分析方法的局限，在疾病与不良事件风险预测、临床辅助诊断、临床辅助治疗、精准医学研究等方面创立了新的应用模式，把基于数据的医学创新带入到更广阔的领域。顺应这一趋势，解放军总医院在大数据领域进行了系统化应用研究，建立了集中的数据资源库，开展了 20 多项数据分析应用，取得了良好的应用效果。由于医学大数据应用刚刚兴起，其发展面临诸多挑战，临床人员大数据思维欠缺、数据质量基础薄弱、开发利用能力不足、医学数据处理分析技术欠成熟的问题有待进一步解决。

关键词：医疗大数据；机器学习；医学创新
中图分类号：R-1 **文献标志码：**A **文章编号：**2095-5227(2019)08-0705-04 **DOI：**10.3969/j.issn.2095-5227.2019.08.001
网络出版时间：2019-8-19 09:07 **网络出版地址：**http://kns.cnki.net/kcms/detail/10.1117.R.20190819.0907.006.html
引用本文：薛万国，应俊 . 大数据时代的医学创新与现实挑战［J］. 解放军医学院学报，2019，40（8）:705-708.

Medical innovation and realistic challenges in the age of big data

XUE Wanguo, YING Jun
Medical Big Data Center, Chinese PLA General Hospital, Beijing 100853, China
The first author: XUE Wanguo. Email: xuewanguo@sina.com

Abstract: In the age of big data, the data environment and analytical techniques that medical research relies on have changed a lot. New predictive analysis and prescriptive analysis, represented by machine learning and deep learning, have overcome the weaknesses of traditional analytical methods. They have been applied in the disease and adverse event risk prediction, clinical auxiliary diagnosis, clinical adjuvant therapy, and precision medical research. Such new application models also bring data-based medical innovation into a broader field. Responding to these trends, Chinese PLA General Hospital has launched a systematic research program in the field of big data, and a centralized data repository has been constructed and more than 20 clinical data analysis tasks are conducted. Most of these applications have achieved good results. However, because the medical big data application is still in its infancy, its development faces some challenges. Problems such as the lack of big data thinking among clinical staff, imperfect data quality, insufficient data managing capability, and inadequate medical data processing and analysis techniques need to be addressed in the future.

Keywords: medical big data; machine learning; medical innovation
Cited as: Xue WG, Ying J. Medical innovation and realistic challenges in the age of big data［J］. Acad J Chin PLA Med Sch, 2019, 40（8）: 705-708.

收稿日期：2019-04-11
基金项目：部委级资助项目；解放军总医院医疗大数据研发项目 (2018MBD-003)
作者简介：薛万国，男，硕士，高级工程师，解放军总医院医学信息情报所副所长，医疗大数据中心主任。研究方向：医疗信息化和医疗大数据应用。Email: xuewanguo@sina.com

医学科技的进步和医院信息化的持续发展推动临床医疗进入了数字化时代。一方面，大容量医学影像检查、微型化的生命体征实时监测等数字化医疗设备广泛应用，以基因数据为代表的新的生命健康数据类型不断出现；另一方面，以电子病历为重点的医院信息化应用持续发展，信息系统覆盖范围越来越广，所采集的医疗数据越来越全面。在这样的背景下，医院存储积累的电子化病例资料越来越多，数据量越来越大，与互联网应用发展进入大数据时代的趋势相呼应，医学也迈入了大数据时代[1]。在大数据发展背景下，医学研究的环境和技术手段发生了很大变化。一方面，医院拥有了“全量”的病例数据，传统的抽样研究得以进入到真实世界研究，拓宽了临床研究的种类，从而能够获得更为多元的医学证据；另一方面，以机器学习、深度学习为代表的大数据分析挖掘技术突破了传统的统计分析方法局限，极大地拓展了临床研究与临床应用的模式。大数据为医学创新发展注入了新的动能，在应用上开辟了新的模式与空间。

1 大数据的分析技术创新

大数据应用开发的核心要义是分析，是围绕特定问题的数据分析技术应用。按照数据分析的目的和分析的深度层级，把大数据分析划分为描述型分析、诊断型分析、预测型分析和指导型分析4种类型。描述型分析是对已发生事实的统计汇总，形成各类描述指标；诊断型分析是对已发生事实的相关因素分析，得出与结果相关的因素及规律；预测型分析是在分析既往数据内在规律的基础上，预测发生某一事件的可能性；指导型分析是在预测未来可能发生的事件基础上，进一步分析给出处置方案。

在既往的医学研究中，对病例数据的分析主要集中在描述型分析和诊断型分析两个层面，典型应用包括病例分布、疗效比较、生存分析、疾病相关因素分析等，所采用的分析技术主要是聚合统计、数理统计、假设检验、回归分析等。随着大数据和数据挖掘技术的发展，机器学习算法获得了充分开发和广泛应用，决策树、随机森林、支持向量机、k- 最近邻、神经网络、逻辑回归、k- 均值等分类、回归、聚类算法把数据分析方法推进到预测型分析层次[2]。特别是由神经网络进一步发展而来的深度学习算法，依赖高性能的计算能力，通过原始数据的训练，可以不再依赖特征的人工选择分析，在大数据的场景下大大提高了预测的准确性，为各类人工智能应用奠定了基础。以机器学习、深度学习方法为代表的大数据分析技术极大地突破了传统的医学研究对疾病及其诊治规律的分析方法局限，能够在弱相关的特征中发现和掌握数据的内在规律，把传统的假设研究进一步发展为探索式研究。新的数据分析技术除了增强传统研究中的描述性及相关性分析能力以外，基于大数据的预测型分析和指导型分析的结果能够直接指导临床实践，通过与业务信息系统的结合，能够多方位地提升临床诊疗的智能化水平，从而把基于数据的医学创新带入到更广阔的领域。

除了数据分析技术外，医疗大数据的处理技术也在持续进步。在临床病例数据中，记录症状、病史、观察记录等重要信息的非结构化文本数据占据了重要位置。目前，通过医学自然语言处理技术已经可以在很大程度上提取出所需的结构化特征，为医学数据的规范化整理提供了有力支持[3-4]。此外，组学测序数据作为内容及结构特殊的一类大数据，其拼接、比对、变异检测以及其他多组学数据的处理分析技术日臻成熟，为精准医学研究的开展奠定了基础。

2 大数据的医学应用创新

医疗大数据除了将传统的抽样研究发展为全样本真实世界研究外，还结合大数据分析技术的新特点，在临床实践中开辟了新的应用模式。

1) 疾病风险预测：疾病的发生发展有一定的演进过程，通常与个体先天因素、个人生活方式、外部环境有密切关联。通过病例大数据分析，发现疾病关联因素并建立疾病预测模型，结合个人当前健康相关状态，可以提前预测罹患某一疾病的风险，如糖尿病风险预测、糖尿病患者发生视网膜病变的风险预测、心衰风险预测[5-7]等。同样，对于疾病患者发生重要的不良事件，也可以通过相关因素建立预测模型，如手术患者发生静脉血栓的风险预测、ICU重症患者死亡风险预测[8]等。通过风险预测，可以及时提醒医护人员对高风险患者提早采取对应的干预措施。

2) 临床辅助诊断：通过对医学文献、病例数据的梳理，建立疾病与症状、检查检验结果、生命体征等的关联，构造出疾病知识图谱。在知识图谱的指导下，可以根据患者的问诊和医学检查，自动给出可能的诊断并指导进一步的检查。在疑难疾病的鉴别诊断方面，也可以通过大数据分析，筛选用于鉴别诊断的相关因素，建立复杂因素与

疾病鉴别诊断预测模型，辅助医生进行相似疾病的鉴别[9]。以影像辅助诊断为代表的医学人工智能呈现出雨后春笋般的发展势头[10]。通过大规模的人工训练，对于肺结节、视网膜病变、肿瘤病理等图像可以做到高精准度识别，效率高且不容易漏诊。临床辅助诊断的应用对于提高诊断质量和诊断效率，特别是对于提高基层医疗机构诊断水平具有重要价值。

3) 临床辅助治疗：同一种疾病，因患者的个体情况不同应采取不同的个性化治疗方案。通过对包含随访数据在内的大样本数据的分析，可以建立不同类型患者的治疗路径组并找出最佳路径，从而为患者提供最适宜的治疗方案。这是在临床指南一般性指导原则之外，根据真实世界数据做出的“第二意见”。此外，通过大量病例数据的学习，建立病例特征分类模型，对于新接诊病例，根据其症状、客观检查等特征，可以在既有病例数据库中自动匹配和推荐相似病例，为新病例的诊断治疗提供参照。

4) 精准医学研究：通过疾病人群与健康人群的基因数据分析，发现基因突变位点和差异基因，找出与疾病相关的基因[11]。通过融合患者的临床表型与生命组学数据，对同一疾病人群进行分类或聚类，细化疾病分型，为疾病的精准治疗提供支持。在疾病诊断方面，针对单一生物标记物特异性不高的问题，通过对疾病人群和健康人群临床与生物特征的聚类分析，可以在多维度体系下，寻找能够更加准确分类的组合特征，从而弥补单一特征的不足，为疾病的诊断鉴别找出新的标记物组合。

3 我院医疗大数据平台建设与应用

为了适应大数据发展的大势，满足医院创新发展对医疗数据的开发利用需求，我院于2016年成立了医疗大数据中心，开始系统化地进行医疗大数据的应用研究。我们整合了我院信息系统中各类数据资源，形成了包含300多万住院患者、4 000多万门诊患者医疗记录的临床数据资源库。面向临床研究开发建设了系列化的医疗大数据利用工具，如医疗大数据检索系统、临床科研数据库系统、病历文本结构化系统等。上述工具实现了把病例数据推送给一线医护人员的目的，解决了过去医护人员临床研究访问使用病例数据难的问题。

两年多来，医疗大数据中心围绕临床问题，组织理工医合作团队，开展了20多项紧贴临床需求的数据分析应用。在病例描述性分析方面，开展了住院患者肿瘤疾病谱、肿瘤治疗方式分布、老年共病、胃癌患者生存分析等研究；在疾病与不良事件风险预测方面，开展了急性胸痛患者疾病鉴别诊断预测分析、糖尿病患者视网膜病变风险预测分析、PCI术后不良事件风险预测分析等研究；在辅助治疗方案选择方面，开展了糖尿病患者用药推荐、乳腺癌患者手术方式选择、手术患者红细胞输注量预测分析等研究；在医学影像人工智能方面，开展了肝癌影像辅助诊断及手术评估系统研发、皮肤黑色素瘤辅助诊断系统研发；在精准医学方面，开展了急性髓系白血病预后不良相关基因、冠心病导致心衰发生过程的分子机制、消化道微生物菌群年龄特征等研究。大部分研究项目取得了很好的分析效果，一些项目研究成果成功转化到临床应用，提高了信息系统的智能化程度，实现了大数据研究从临床中来到临床中去的闭环。

4 医疗大数据应用面临的挑战

总体上，医疗大数据的应用才刚刚兴起，从技术到应用都还有待进一步发展，当前也面临着一些困难和挑战[12]。

1) 临床人员大数据思维还较为欠缺：基于数据思维提出临床问题是大数据应用创新的前提。当前医疗大数据发展的呼声很高，一些临床人员也重视数据的积累，在数据的整理、清洗和平台的建设上投入了不少精力，但在利用数据解决什么问题上却思路不多。一些数据分析也仅限于描述性统计，创新性研究案例较少。打开大数据应用创新的钥匙在于激发临床人员的问题与思路。这既需要大数据知识的普及，也需要典型创新应用的示范引领。

2) 医疗数据质量基础还较为薄弱：完整和准确的病例数据是高质量大数据分析的基础。遗憾的是，虽然近年来我们的医疗信息化建设有了长足进步，但数据的完整性还有不少欠缺。首先是随访数据普遍欠缺，导致临床研究缺乏结局对照，极大地限制了数据的可用性；其次，由于医疗业务的复杂多样性，信息系统覆盖不完善或者集成度不够，导致一些对临床研究非常重要的专科数据缺失；在数据准确性方面，由于许多医疗记录为自由文本，加上记录上的随意性，导致部分数据质量不高；另外，由于数据共享难的问题尚待破解，多中心数据研究就更为困难。解决以上问题，既需要持续地有针对性地发展完善信息系统，

也需要加强信息化应用管理和基础医疗质量管理。

3) 开发利用大数据的能力不足：开展大数据应用研究涉及计算机、统计分析、生物信息、生物医学工程等多个学科，需要建立多学科人才的合作团队。但传统上，医院信息中心的主要工作是信息系统的建设和运行维护，在人才方面以计算机专业为主，缺乏数据分析方面的专业人才。在数据服务方面，没有建立规范化的数据服务流程和模式，数据服务水平不高。提升医院的大数据利用能力，需要强化信息中心的数据服务职能，重构信息中心的人才队伍，建立团队协作机制。

4) 医疗大数据处理分析技术还欠成熟：医疗数据中文本数据、影像数据、波形数据等类型繁多，是各行业数据中最为复杂的一类。虽然这些非结构化数据的处理技术在不断发展，但病历文本特征提取、影像特征提取、组学数据特征提取等还缺乏成熟易用的工具，许多情况下仍需要人工通过较为初级的工具去完成。这方面还需要产学研共同努力，重点发展。

5 结语

大数据为医学的创新发展注入了新动力。随着数据资源的积累、数据分析技术的进步和临床人员的广泛参与，大数据在医学中的应用模式将得到更加充分的展现和更大程度的推广，这不但可以提高临床工作质量与效率，而且将创新传统医学工作模式，使得人工智能、智慧医疗、人机协同等未来医学模式成为可能。

参考文献

1 邹北骥. 大数据分析及其在医疗领域中的应用[J]. 计算机教育，2014，(7)：24-29.
2 秦文哲. 大数据背景下医学数据挖掘的研究进展及应用[J]. 中国胸心血管外科临床杂志，2016，23(1)：55-60.
3 李国垒，陈先来，夏冬. 面向临床决策的电子病历系统概述[J]. 中国数字医学，2014，9(12)：30-32.
4 Rajkomar A，Oren E，Chen K，et al. Scalable and accurate deep learning with electronic health records[J]. NPJ Digital Medicine，2018，1(1)：18.
5 王喜丹，王晓丹，梁丽. 基于深度学习模型在2型糖尿病患病风险预测中的应用[J]. 临床医药文献电子杂志，2017，4(84)：16460-16461.
6 曹文哲，应俊，陈广飞，等. 基于Logistic回归和随机森林算法的2型糖尿病并发视网膜病变风险预测及对比研究[J]. 中国医疗设备，2016，31(3)：33-38.
7 苏枫，张少衡，陈楠楠，等. 基于机器学习分类判断算法构建心力衰竭疾病分期模型[J]. 中国组织工程研究，2014，18(49)：7938-7942.
8 王力红，赵霞，张京利，等. 重症监护病房医院感染预警模型的建立[J]. 中华医院感染学杂志，2010，20(21)：3368-3370.
9 Liang H，Tsui BY，Ni H，et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence[J]. Nat Med，2019，25(3)：433-438.
10 许晶晶，谭延斌，张敏鸣. 影像学在肿瘤精准医疗时代的机遇和挑战[J]. 浙江大学学报(医学版)，2017，46(5)：455-461.
11 赵学彤，杨亚东，渠鸿竹，等. 组学时代下机器学习方法在临床决策支持中的应用[J]. 遗传，2018，40(9)：693-703.
12 张振，周毅，杜守洪，等. 医疗大数据及其面临的机遇与挑战[J]. 医学信息学杂志，2014，35(6)：1-8.