综　　述

大数据技术在药学领域应用的研究进展

黄　婧, 李金慈, 苏　华综述, 郑　均审校

【摘要】 随着科学技术的发展, 药学领域产生庞大且复杂的数据集, 需要使用大数据技术对其进行分析, 使得过程更加高效快捷。为了更好地了解大数据技术在药学领域的应用, 文章介绍了大数据相关技术及其目前在新药研发、药动学、毒理学、中药学、临床药学中的应用, 并对目前仍存在的问题及发展趋势进行分析, 为后期深入应用提供参考。

【关键词】 大数据技术; 药学领域; 新药研发; 药动学; 中药学; 临床药学

【中图分类号】 R91　　【文献标志码】 A　　【文章编号】 1672-271X(2023)01-0080-05

【DOI】 10.3969/j.issn.1672-271X.2023.01.016

Research progress on application of big data technology in pharmacy

HUANG Jing[1], LI Jinci[1], SU Hua[1] *reviewing*, ZHENG Jun[2] *checking*

(1. *Department of Preparation Division*, 2. *Medical Security Centre*, *General Hospital of Eastern Theater Command*, *PLA*, *Nanjing* 210002, *Jiangsu*, *China*)

【**Abstract**】 With the development of science and technology, the pharmaceutical field produces huge and complex data sets, which need to be analyzed by using big data technology to efficiently and quickly make the process. In order to better understand the application of big data technology in pharmacy, this review introduces the technology of big data and its current application in the research and development of new drugs, pharmacokinetics, toxicology, traditional Chinese medicine and clinical pharmacy. We analyze the existing problems and development trends, so as to provide reference for further application in the future.

【**Key words**】 big data; pharmacy; development of new drugs; pharmacokinetics; traditional Chinese medicine; clinical pharmacy

0　引　　言

大数据时代科学研究是一个大科学、大需求、大数据、大计算、大发现的过程, 中国科学院早在 1982 年就正式提出科学数据库及其应用系统项目建设[1]。《Nature》于 2013 年出版了“大数据”专刊, 吸收来自互联网技术、生物医学、超级计算、环境科学等多个科技方面大数据的应用[2]。随着研究数据的大幅增长, 围绕药学领域的数据库应运而生, 大数据技术在药学领域的应用相关研究越来越多, 切实解决了很多问题, 并且高效快捷, 已成为学科发展和技术进步的迫切需要。本文就目前大数据技术在药学领域的应用现状进行综述, 为进一步探索大数据技术在药学领域应用提供参考。

1　大数据技术简介

随着信息技术的发展和数据爆炸式增长, 产生庞大而复杂的数据集, 由此产生的以数据为中心的环境要求我们去获取、整合和分析大数据, 以破解复杂的科学问题。这种揭示有意义潜在发现的复杂数据挖掘过程, 被称为大数据分析[3]。大数据技术的出现已经彻底改变了药学领域研究的过程与策略, 使得研究进度加快, 效率更高, 减少研究经费。

大数据分析是处理复杂数据很好的解决方案, 提高了决策能力, 其特点主要是 5V: 数量巨大(Vol-

基金项目: 全军医疗机构制剂标准提高科研专项课题重点项目(14ZJZ08)

作者单位: 210002　南京, 东部战区总医院药剂科(黄　婧、李金慈、苏　华), 医疗保障中心(郑　均)

通信作者: 郑　均, E-mail: zj-dongzong@163.com

ume)，数据源多变(Variety)，数据处理速度快(Velocity)，数据真实(Veracity)、数据提供有效价值(Value)[4-7]。大数据分析在药学领域主要是指从大量的数据集成和复杂的异构数据，如各种数据库(Drugbank、ZINC、CPDB 等)，各种组学(基因组学、蛋白质组学、代谢组学等)，获取有价值的信息，为后续研究提供快速解决办法。数据挖掘、统计分析、机器学习、神经网络这些技术是在大数据集基础上做出更好、更快的决策[8]。大数据技术在药学领域的应用越来越受到关注。

2 大数据技术在药学领域的应用

2.1 大数据助力新药开发 大数据可应用在新药创制的不同环节，预测药品的安全性、有效性、不良反应等对研发成败起到关键作用的药物属性，包括虚拟筛选苗头化合物、药物分子设计、新药合成路线设计、药物有效性及安全性预测等。能切实减少人力、物力、时间等研发投入，从而降低药品研发成本和风险，缩短医药创新成果转化的过程[9]。用于新药发现的大数据可分为不同类别的数据库存储，包括化合物数据库(如 PubChem，ChEMBL)、药物信息数据库(如 DrugBank，e-Drug3D)、药物靶标数据库，包括基因组或蛋白质组学数据库(如 Binding DB，SuperTarget)、分析筛选代谢和疗效研究的数据库(如 HMDB、TTD)等[10]。

Fang 等[11]通过对 Binding DB 数据库挖掘得到了有丁酰胆碱酯酶抑制活性的化合物，并结合本实验室筛选得到的对丁酰胆碱酯酶无抑制活性的化合物，建立支持向量机(SVM)及朴素贝叶斯(NB)模型，然后对化合物样品库进行虚拟筛选，最后将模型预测的 30 个抑制剂进行生物测定，结果有 10 个化合物具有很好的丁酰胆碱酯酶抑制活性，大大提高了筛选效率。

Kumar 等[12]利用 ZINC 和 DrugBank 数据库收集对白血病融合基因(BCR-ABL)有抑制作用的药物。将筛选药物与已有报道药物对比，研究筛选药物与 ABL 酪氨酸激酶活性位点的对比对接分析。对筛选出的最佳相互作用配合物进行 50 ns 分子动力学模拟，验证系统稳定性，用计算机模拟对筛选出的抑制剂进一步验证和分析的药代动力学性质和一系列吸收、分布、代谢、排泄、毒性(ADMET)参数分析，以获得其药物性质。通过此药物设计，筛选出多个分子可作为抗慢性骨髓白血病的潜在先导物，对于研究治疗抗慢性骨髓白血病的新药中有极大的影响。

王哲[13]基于 2P2IDB 数据库，分别对现有蛋白-蛋白相互作用(PPI)小分子抑制剂的结构特征和 PPI 靶标结合的特点进行统计分析，构建了 PPI 小分子抑制剂对接构象排序性能测试数据集。基于分子对接、分子力学/泊松-波尔兹曼(广义波恩)表面积[MM/PB(GB)SA]和伞形采样方法对一系列嘧啶酮类泛素特异性蛋白酶 7(USP7)小分子抑制剂的结合强度进行了预测，发现了强结合抑制剂和弱结合抑制剂的解离路径间存在较大差异。基于三维形状的相似性搜索等方法针对 USP7 的非催化位点开展了层级式和反馈式的虚拟筛选，购买并测试了近 200 个化合物，最终发现了数个结构新颖、活性较强的苗头化合物。

2.2 大数据技术在药动学研究中的应用 生理药代动力学(PBPK)建模一直用于预测药代动力学参数，集成了大量的药物特异性数据、参数和物种解剖生理学数据，这些机体的内在特征性参数来源于生物大数据[14]。PBPK 模型是“基于机理”的模型，可进行体外到体内、动物到人体以及普通人群到特殊人群的合理外推[15]，实现跨种属跨人群障碍的药动学模拟[16]。

李正[17]利用 GastroPlus® 软件，基于机体生理学性质、生理模型结构和药物性质，搭建并验证阿替美唑的大鼠 PBPK 模型。应用大鼠 PBPK 分布模型定义人体的分布模型，运用体外酶动力学参数定义清除模型，建立人体 PBPK 模型，并用文献查阅美国人静脉注射阿替美唑的药动学数据进行模型验证。进而预测中国人群肌肉注射阿替美唑后体内血浆和靶组织的分布。结果表明，肌肉注射阿替美唑浓度范围在 500(mg/人)以内在人体内以及脑组织中呈线性暴露，超过 1000(mg/人)才会出现非线性暴露。为后期的临床实验提供有效的数据支持。

Kohlmann 等[18]研究儿童、婴儿、新生儿的年龄差异对卡马西平口服吸收的影响。在 GastroPlus® 软件中建立了口服吸收模型，评估成人 PBPK 模型，利用临床数据外推到更小年龄，并对不确定的模型参数进行敏感性分析。结果显示卡马西平吸收溶解受溶解度、颗粒半径、小肠转运时间等因素的影响，这些因素与患儿年龄及卡马西平剂量有关。

此研究有利于更好的了解儿童患者的药物口服吸收。

2.3 大数据技术在毒理学研究的应用 药物开发过程中,安全性一直是最重要的问题,包括各种毒副作用和不良反应。目前构建毒性预测模型一般包含四个步骤:数据收集、数据描述、模型构建、模型评估。数据收集是通过数据库收集化合物结构、安全性、靶点、信号通路等信息;数据描述是用化学结构、物理化学信息或拓扑特征计算出分子描述符来描述或用二进制串表示的分子指纹;通过机器学习的方法建立预测模型,可以是单个模型,也可以是多个模型集成;最后通过模型验证评估其准确性。

Zhang 等[19]从 CPDB 数据库中收集 1003 个化合物,包括 494 个致癌物及 509 个非致癌物作为训练数据,用于建立和验证预测模型。以分子量、溶解性、氢键受体数等参数生成 12 种类型的分子指纹。利用 7 种类型的分子指纹和 3 种类型机器学习方法,开发了 3 种新的集成分类模型,即集成支持向量机、集成随机森林和集成 XGBoost 算法,来预测化学品的致癌性。将此集成模型用于 DrugBank 数据库发现潜在致癌物,结果发现此模型有利于潜在致癌物发现。

2.4 大数据在中药学领域的应用 中药复方往往由多种药味组成,每种药味又含有多种成分,因此研究较为复杂且困难。随着大数据理念的深入,越来越多研究大数据在中药领域的应用,形成了系统化、集成化的大数据应用平台,为中药事业开拓发展提供支持。

2.4.1 应用于中药活性成分筛选 中药系统药理学数据库分析平台(TCMSP)是包括化学物质、靶点和药物靶点网络,以及相关的药物靶点网络,以及涉及相关的药物靶点网络,以及涉及口服生物利用度、药物相似度、肠上皮通透性、血脑屏障、水溶性等天然化合物的药代动力学特性的数据库。多数学者利用 TCMSP 平台,采用计算机虚拟筛选技术实现分子对接,筛选有效活性成分[20-21]。李婧等[22]采用 TCMSP、CNKI、PubMed 数据库整理搜集候选中药主要活性成分及有抗病毒活性的成分,对 SARS-CoV-23CL 水解酶(Mpro)蛋白采用 AutoDock Vina 进行分子对接。共得到 11 个高频使用抗病毒待研究中药,及 469 个候选活性成分。从传统抗病毒中药中筛选出潜在的抗新型冠状病毒(SARS-CoV-2)中药单体,为抗 SARS-CoV-2 药物研究及处方筛选提供参考。

2.4.2 应用于中药制剂生产 中药制剂工业化生产产生一系列生产记录和检验报告等海量数据,从数据中挖掘质量传递规律,找出影响质量的关键工艺参数,实现产品质量的追溯与控制,促进中药制造过程精益操作、优化和持续改进[23]。杜慧等[24]收集 2017-2018 年热毒宁注射剂的历史生产数据 259 批,共计 829 318 数据点,以金青醇浓缩制得浸膏为响应变量,通过数据清洗和特征提取。采用皮尔逊(Pearson)相关分析和灰色关联度分析进行综合决策,从特征变量中筛选出潜在关键工艺参数。从全局数据出发,采用大数据分析的方法有效提高数据的价值密度,筛选得到的关键工艺参数有助于解析金青醇沉生产过程的质量传递规律。

2.5 大数据技术在临床药学中的应用

2.5.1 临床合理用药 2007 年,美国药剂师协会(ASHP)发布文件将数据信息、药学知识、自动化技术进行整合,以更好的服务于临床合理用药,改善患者用药安全[25]。各国都以将健康医疗大数据定位为国家战略,而临床合理用药则是医疗大数据挖掘分析的关注点之一。这不仅是医院本身的用药监测专业化工具,更能够切实降低患者用药风险、提高用药合理性、有效性和经济性。①应用于个体化给药:个体化给药是指根据患者个体特征,结合药动学-药效学原理及临床药物治疗指南,制定个体优化的治疗方案[26]。群体药动学(PPK)结合贝叶斯估算的最大后验贝叶斯法(MAPB)是目前公认的最佳剂量计算方法[27-28]。高玉成等[29]通过系统检索,收集中国人群的万古霉素群体药动学特征参数,,结合 R 语言贝叶斯分类算法包的最大后验贝叶斯算法, 研制了万古霉素的个体化给药决策辅助系统“SmartDose”。该系统可针对普通成人以及新生儿、老年人、神经外科患者等特殊人群, 制定个体化的万古霉素给药方案。系统功能包括制定初始方案、根据治疗药物监测结果调整方案, 以及自定义用药方案等。该系统适用面广, 为万古霉素的个体化用药提供了有力的工具。②应用于审方:合理用药是指安全、有效、经济地使用药物。随着医院信息化深入,以临床用药为基础,收集数据构建药物信息平台,通过制定用药审核规则行为,进行规范审核和风险提示。洪灵鸿[30]以两家三甲医院

儿科处方大数据建立用药知识库,利用 Apache Spark 分布式计算框架提取不同药品的用药剂量历史记录,以及其对应的患儿生理信息、适应症类型、合并用药种类等相关特征因素。基于机器学习库 MLlib 建立不同药品的用药剂量关于上述特征因素的混合预测模型,在此基础上对新的处方用药剂量进行风险评价;基于图模型计算库 GraphX 建立不同药品之间联用的概率图模型,在此基础上利用子图搜索算法,对新的处方中的药品联用进行风险评价,从而为药师审方提供决策依据。

2.5.2 应用于药品不良反应的预测 对于药物不良反应的研究通常进行体内、体外试验,近年来还在开发一些体外模型,如器官芯片,来降低研究成本[31-32],但这些方法仍然需要大量的时间与成本。比较于这些方法,计算机方法能快速、低成本、相对准确预测出药品不良反应[33]。

药物性肝损伤是最常见的药物性不良事件之一,可导致急性肝衰竭等危及生命的情况。药物性肝损伤临床表现复杂且特殊,需要开发新的预测方法来进行评估。Zhu 等[34]利用上市后数据构建了一个肝毒性的数据库,利用定量关系-活性结构(QSAR)建模。选择了 37 个与肝毒性相关的首选词从数据库中提取数据,提取出 2029 种可建模的药物,包含 13 555 个药物-不良反应组合。在已有文献基础上,利用阳性及阴性药物对模型进行校准,从而优化模型的预测性能。

3 目前存在问题和发展趋势

3.1 缺乏大数据分析专业人才 普通的药学工作者缺乏大数据相关背景,对数学、统计知识的掌握不够成熟,算法建模、软件设计等技能缺乏,使得很多分析工作无法顺利开展,这些都会阻碍大数据技术在药学实践中的应用。药学领域需要更多的数据分析人才加入,因此,未来多学科交叉人才的培养是发展药学学科的新需求。

3.2 从海量数据中获取有效数据 在收集数据的同时也产生大量无用、关联性低、价值密度低的数据,而数据的价值、变异性和准确性对于数据的实际应用来说,更值得考虑[35]。从海量数据中如何筛选出真实、有效的信息以得到正确的计算是对于分析人士很大的挑战。另外,药物研究中的大量珍贵数据处于私有状态,能公开获取的数据不足,这也是药学领域进入大数据时代的一大挑战。

3.3 未来发展趋势 随着大数据时代的到来,科学技术的高速发展,大数据在药学领域的应用也越来越深入,也在不断产生的新技术、新方法,为药学研究提供有效的分析方案,并切实缩短研究时间,节约研究经费。大数据分析技术在药学领域发挥的作用也越来越显著。随着学科发展,技术革新,未来的应用会更广泛,更深入,为药学领域带来新的发展。

【参考文献】

[1] 黎建辉,沈志宏,孟小峰. 科学大数据管理:概念、技术与系统[J]. 计算机研究与发展,2017,54(2):235-247.

[2] Nature [EB/OL]. [2014-08-23]. http://www.nature.com/news/specials/bigdata/index.html.

[3] Uthayasankar S, Muhammad MK, Zahir I, *et al*. Critical analysis of Big Data challenges and analytical methods[J]. J Business Res, 2017,70: 263-286.

[4] Wang Y, Kung L, Wang WYC, et,al. An integrated Big Data analytics-enabled transformation model: application to health care [J]. Inf Manag,2018,55(1):64-79.

[5] Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations [J]. Technol Forecast Soc Change, 2018,126:3-13.

[6] Oppitz M, Tomsu P. Inventing the cloud century[M]. Cham: Springer, 2018.

[7] Yang C, Yu M, Hu F, *et al*. Utilizing cloud computing to address big geospatial data challenges[J]. Comput Environ Urban Syst, 2017,61:120-128.

[8] 敖 翼,濮 润,卢 姗,等. 我国新药创制的模式选择与发展思考[J]. 中国新药杂志,2020,29(2):136-142.

[9] Rahmani AM, Azhir E, Ali S, *et al*. Artificial intelligence approaches andmechanisms for big data analytics: asystematic study [J]. PeerJ Comput Sci,2021,7(2):e488.

[10] Tripathi MK, Nath A, Singh TP, *et al*. Evolving scenario of big data and Artifcial Intelligence (AI) in drug discovery[J]. Mol Divers,2021,25(3):1439-1460.

[11] Fang JS, Yang RY, Gao L, *et al*.Predictions of BuChE inhibitors using support vector machine (SVM) and naive Bayesian classification techniques in drug discovery[J]. J Chem Inf Model,2013, 53(11):3009-3020.

[12] Kumar H, Raj U, Gupta S, *et al*. In- silico identification of inhibitors against mutated BCR- ABL protein of Chronic Myeloid Leukemia: A Virtual Screening and Molecular Dynamics Simulation study [J]. J Biomol Struct Dyn, 2016,34(10):2171- 2183.

[13] 王 哲. 基于分子对接的虚拟筛选方法的评测、优化和应用[D]. 杭州:浙江大学,2019.

[14] Ivan Nestorov. Whole Body Pharmacokinetic Models[J]. Clin

Pharmacokinet, 2003, 42:883-908.

[15] Rioux N, Waters NJ. Physiologically-Based Pharmacokinetic Modeling in Pediatric Oncology Drug Development[J]. Drug Metab Dispos, 2016, 44(7): 934-943.

[16] Dinh JC, Pearce RE, Van Haandel L, *et al*. Characterization of Atomoxetine Biotransformation and Implications for Development of PBPK Models for Dose Individualization in Children[J]. Drug Metab Dispos, 2016, 44(7): 1070-1079.

[17] 李 正. 探索新药阿替美唑跨种属障碍人体药动学仿真[D]. 北京:军事科学院,2019.

[18] Kohlmann P, Stillhart C, Kuentz M, *et al*. Investigating Oral Absorption of Carbamazepine in Pediatric Populations[J]. AAPS J, 2017, 19(6): 1864-1877.

[19] Zhang L, Ai HX, Chen W, *et al*. CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods[J]. Sci Rep, 2017,7(1):2118.

[20] 马青云,刘 辰,杜海涛,等. 基于高通量分子对接虚拟筛选 SARS-CoV-2 3CL 水解酶中药小分子抑制剂及抗新型冠状病毒肺炎(COVID-19)的中药及其复方预测[J]. 中草药,2020,51(6):1397-1405.

[21] 秦健峰,郝二伟,梁跃辉,等. 基于文献挖掘与分子对接技术的瑶药“十八钻”抗新型冠状病毒活性成分筛选[J]. 中草药,2020,51(8):2024-2034.

[22] 李 婧,马小兵,沈 杰,等. 基于文献挖掘与分子对接技术的抗新型冠状病毒中药活性成分筛选[J]. 中草药,2020,51(4):845-850.

[23] 徐 冰,史新元,罗 赣,等. 中药工业大数据关键技术与应用[J]. 中国中药杂志,2020,45(2):221-232.

[24] 杜 慧,徐 冰,徐芳芳,等. 大数据驱动的热毒宁注射液金青醇沉关键工艺参数辨识研究[J]. 中国中药杂志,2020,45(2):233-241.

[25] Ma C, Wong H, Chu C, *et al*. Big data in pharmacy practice: current use, challenges, and the future[J]. Integr Pharm Res Pract, 2015:4 91-99.

[26] 凌 静,焦 正,钟明康. 目标浓度干预概况及研究进展[J]. 中国药学杂志,2013,48(16):1337-1342.

[27] Zhao CY, Jiao Z, Mao JJ, *et al*. External evaluation of published population pharmacokinetic models of tacrolimus in adult renal transplant recipients [J]. Br J Clin Pharmacol, 2016, 81: 891-907.

[28] Mould DR, Dubinsky MC. Dashboard systems: pharmacokinetic/pharmacodynamic mediated dose optimization for monoclonal antibodies [J]. J Clin Pharmacol, 2015, 55 Suppl 3: S51-S59.

[29] 高玉成,焦 正,黄 虹,等. 万古霉素个体化给药决策支持系统的研制[J]. 药学学报,2018,53(1):104-110.

[30] 洪灵鸿. 大数据技术在儿科临床合理用药中的应用初探[D]. 杭州:浙江大学,2018.

[31] Huh D, Matthews BD, Mammoto A, *et al*. Reconstituting organ-level lung functions on a chip[J]. Science, 2010, 328(5986): 1662-1668.

[32] Huh D, Hamilton GA, Ingber DE.From 3D cell culture to organs-on-chips[J]. Trends Cell Biol, 2011, 21(12):745-754.

[33] Segall MD, Barber C. Addressing toxicity risk when designing and selecting compounds in early drug discovery[J]. Drug Discov Today, 2014,19(5):688-693.

[34] Zhu X, Kruhlak N. Construction and analysis of a human hepatotoxicity database suitable for QSAR modeling using post-market safety data[J]. Toxicology,2014,321:62-72.

[35] Andreu-Perez J, Poon CC, Merrifield RD, *et al*. Big data for health[J]. IEEE J Biomed Health Inform, 2015, 19(4): 1193-1208.

(**收稿日期**:2022-10-18; **修回日期**:2023-01-11)

(**责任编辑**:刘玉巧; **英文编辑**:朱一超)