·述评·

大型队列研究中的数据科学

余灿清 李立明
北京大学公共卫生学院 100191
通信作者:李立明, Email:lmleeph@vip.163.com

【摘要】 大型队列研究作为生物医学研究的宝贵资源,在揭示疾病的病因、发病机制,改善疾病预后,减轻疾病负担等方面显示了巨大的作用。数据科学是一门新型的交叉学科,它采用计算机科学和统计学方法与特定专业领域相结合,发现数据背后的规律和知识。两者的结合为我国疾病防治策略和措施的制定提供新的证据。本文介绍了数据科学的基本概念,并结合大型队列研究的特点和发展趋势,分析队列研究数据的内容和结构特点,探讨了数据科学在大型人群队列的不同研究阶段的应用和价值,以及数据科学在大型队列研究中的应用前景。

【关键词】 大型队列研究; 数据科学
基金项目:国家重点研发计划(2016YFC0900500,2016YFC0900504)
DOI:10.3760/cma.j.issn.0254-6450.2019.01.001

Data science in large cohort studies

Yu Canqing, Li Liming
School of Public Health, Peking University, Beijing 100191, China
Corresponding author: Li Liming, Email: lmleeph@vip.163.com

【**Abstract**】 Large cohort study gained its popularity in biomedical research and demonstrated its application in exploring disease etiology and pathogenesis, improving the prognosis of disease, as well as reducing the burden of diseases. Data science is an interdisciplinary field that uses scientific methods from computer science and statistics to extract insights or knowledge from data in a specific domain. The results from the combination of the two would provide new evidence for developing the strategies and measures on disease prevention and control. This review included a brief introduction of data science, descriptions on characteristics of large cohort data according to the development of the study design, and application of data science at each stage of a large cohort study, as well as prospected the application of data science in the future large cohort studies.

【**Key words**】 Large cohort study; Data science
Fund programs: National Key Research and Development Program (2016YFC0900500, 2016 YFC0900504)
DOI:10.3760/cma.j.issn.0254-6450.2019.01.001

队列研究(cohort study)被广泛地用于某一特定暴露和(或)多种暴露与一种或多种疾病结局的关联性研究,在验证病因假设方面具有不可替代的地位和作用。随着人们对复杂性疾病病因研究的不断深入,研究内容逐渐从单个环境或遗传因素的发现和验证,发展到更为复杂的基因-基因、基因-环境的交互作用。这种深入、精细的研究需要更大的样本量、更为准确全面的测量。因此,大型人群队列研究被推到了医学研究热点的同时,也将如何收集数据、清理数据、分析和解释数据提高到了前所未有的地位。一时间,大数据、数据挖掘、机器学习等概念陆续登台,直到数据科学(data science)的出现,才给庞杂的数据工作进行全面的梳理。本文从数据科学的基本内容和研究范畴出发,探讨其在大型队列研究中的应用和价值。

一、数据科学

数据科学是采用各种科学方法、处理过程、计算算法、系统体系从各种数据中提取和发现知识的一门交叉学科[1]。通过研究不同数据类型、状态和属性及其变化规律,探讨如何对数据处理和分析,从而揭示自然界和人类行为等现象背后的规律。数据科学采用了数学、统计学、信息科学和计算机科学等多个学科的技术和原理,去分析和理解数据中的实际现象。数据科学与数据库和大数据的区别在于,数据库通过对数据的增删改查等操作和简单报表,积累了大量的基础数据,为数据科学提供了重要的"原

材料”；而大数据及其分析方法，例如机器学习，数据挖掘等，是数据科学的有机组成部分。

从事数据科学工作需要掌握多个学科的知识，Geringer描绘了该学科的维恩图（Venn diagram），主张将这些知识归纳为3个方面[2]。首先是数学和统计学知识，包括高等数学、数学分析、概率论与数理统计、多变量分析、机器学习、深度学习，以及非结构化数据的分析等内容；其次是计算机科学知识，包括计算机基础、数据库结构、编程与算法、并行计算与脚本语言等；最后是某特定领域的专业知识，即利用专门学科的知识来指导和制定数据分析和挖掘的策略，发现隐藏在该领域各种数据背后的规律。对于公共卫生来说，专业知识包括生物学、医学和预防医学等学科的相关知识。在开展流行病研究中，特别强调对疾病病因和因果推断的理解，这是指导统计分析和建模的核心，从而梳理和系统化数据科学发现的规律和知识。

2012年，《哈佛商业评论》中称“数据科学家是21世纪最性感的工作”[3]，从此数据科学成为业界的一个热词。实际上，数据科学的方法已经广泛运用在数据密集的全球化商业和金融领域。我们日常生活中应用的例子也不鲜见，比如实时在线广告推送、物流运输线路设计、空中交通管制等。近些年来，随着生物医学、各种组学和影像学技术的发展，产生了丰富、海量的医学数据，给疾病的筛查、诊断、治疗和预防带来新的视角和发展契机。数据科学与医学的深入交叉和融合，将进一步促进肿瘤、心血管疾病和传染病等常见疾病防治的个体化和精准化。

二、大型人群队列研究

队列研究设计是经典的分析性流行病学方法之一，前瞻性地观察和比较不同暴露水平人群发生疾病或健康相关结局的差异。这种研究设计具有同期的对照，满足由因及果的前瞻性时序关系，具有较强的因果检验能力，是目前开展复杂疾病病因、预后及疾病负担的最佳研究设计[4]。

从20世纪开始，随着环境和生活方式等因素日益突出，研究者开始人群队列研究评估这些环境因素以及遗传对健康的影响[5-6]。始建于1948年的美国Framingham队列研究就是其中的成功典范，通过过去70年的不懈努力，对3代研究对象进行追踪研究，累计发表3 698篇科研论文[7]，该队列成为了心脑血管疾病的病因研究平台，确定了心脑血管病的主要危险因素，为疾病防治指南提供了科学证据，也变革了人类的健康观念。随后，欧美一些发达国家陆续建立了一系列长期随访的人群队列，例如，英国男性医生队列研究、英国全国卫生与发展调查（National Survey of Health and Development）、英国出生队列（British Birth Cohort）、英国百万女性研究（the Million Women Study）、英国生物样本库（UK Biobank）、美国护士健康研究（Nurses’ Helath Study），美国军人千禧队列研究（Millennuium Cohort Study）、美国家庭队列研究（Cohort Family Study）和欧洲癌症和营养前瞻性调查（the European Prospective Investigation into Cancer and Nutrition）等。这些队列项目从不同人群、不同角度开展研究，发现和验证了多种暴露因素（尤其是生活方式和环境因素暴露）与主要慢性病之间的关联。

相比之下，我国队列研究起步较晚，规模较小、研究分散、项目执行期短且缺乏长期稳定支持。比较成功的大型人群队列多以常见病为主，例如，1986年和1996年由上海肿瘤研究所分别建立的上海市男性和女性健康队列，1989年开始的中国健康与营养调查队列，1992年建立的11省市区心血管发病前瞻性队列，2004年启动的中国慢性病前瞻性研究和2007年启动的泰州大型人群健康队列等。2016年科技部将“精准医学研究”纳入国家重点研发计划的重点专项，旨在建设大规模人群队列研究，内容涵盖了百万人以上的国家级大型自然人群健康队列、重大疾病专病队列和罕见病的临床队列，建立多层次精准医学知识库体系和安全稳定可操作的生物医学大数据平台[8]。

三、大型人群队列研究与数据科学

1. 大型人群队列研究的数据特点：大型人群队列研究的数据具有大数据“4V”的部分特点。首先是样本量较大，一般从几万到几十万不等，随访时间长达十几年甚至几十年，积累了海量的研究数据，即具有大数据容量大（Volume）的特点。近些年，随着生命科学和现代信息技术的迅速发展，基于队列研究建立生物样本库（Biobank）已成为流行趋势。队列研究进一步整合基因组学、表观组学、蛋白组学、代谢组学等多维度数据，数据量迅速扩大。这些海量、多维度数据的积累将更好地帮助我们理解疾病发生、发展的生物学机制，促进以大型队列为基础的系统流行病学研究的发展，逐个破解威胁人类健康主要疾病病因通路中的黑匣子。

然而，与大数据低价值密度（Value）、大量无意义的冗余、垃圾数据不同。大型人群队列建设是一项长期复杂的系统工作，研究者在研究初期制定严

谨、可行的研究设计方案和实施计划，在现场调查和数据收集过程有严格的质控措施。因此，研究对象具有较好的代表性，研究数据通常具有较高的质量和真实性（Veracity），能更好地解释数据背后的规律和因果关系。

大型队列的数据也具有多样性（Variety）。从近些年的发展趋势来看，队列研究的暴露因素从传统的职业、环境、饮食等因素扩大到遗传、行为、心理、社会、生态等各个方面。随访数据获得方法也从传统的主动调查，发展到与现代医疗、社会保障体系、公共安全体系以及生命统计和疾病监测体系等现有数据平台联网与整合，获取调查对象的各种结局的信息。因此，大型队列研究数据具有足够充分、完整的信息，可以开展具有相当广度和深度的大数据分析和科学研究。

此外，队列研究在实施过程中，自身也产生了一系列辅助化的数据，例如项目管理记录、系统运行日志、仪器设备调度和运行记录，这些也可以作为队列研究数据的有效补充。

2. 大型队列研究的数据结构特点：大型队列研究作为精准医学研究资源平台，除了主动收集高质量的研究数据外，还保持与其他数据互通互联的特性，这也是数据科学与之结合的生发点。随着医学大数据和精准医学的深入结合，队列研究数据逐步实现与生物大数据、临床大数据和健康大数据的整合，包括与多组学数据、电子病历与健康档案数据、医学检验与影像学资料、公共卫生监测、个体智能设备与互联网健康数据等多源异构数据的对接[9]，甚至与地理、气象、交通等数据的互通。

虽然传统队列研究的数据结构多以关系型数据库为主，但大型队列研究收集的数据类型多样，兼有结构化、半结构化和非结构化数据。例如，现场调查的影像、调查录音、检测仪器产生的结果文件和日志等，需要采用数据科学的方法进行提取和标准。同时，由于数据量较大，需要考虑数据存储和检索效率，以及分析利用的便利性，需要更为高级的数据库手段来实现日常管理和科学研究的需要。

与大数据不同的是，大型队列研究通常在实施前具有详细的研究设计、翔实的实施方案和备忘文档。对实施过程中的常规专项工作，研究组通常通过制定内部操作规范（IOP）和标准操作流程（SOP），确保操作的统一性，也便于研究组内传递和共享。同时，大型队列作为优秀的科研资源平台，容易吸引国内外高水平的研究者积极参与，产出高质量的学术论文，并有成熟的技术整理出版，详细介绍数据收集和测量方法、处理过程和规则。这些文档都有利于队列数据的标准化和合作共享。

3. 数据科学在大型队列研究的应用：数据科学的生命周期分为收集（capture）、维护（maintain）、处理（process）、分析（analyze）和交流（communicate）5个阶段，分别与大型队列研究的不同实施环节相结合，产生不同的工作内容[10]。在大型队列研究数据收集阶段，数据科学可以采用计算机技术手段辅助数据获取和录入，提高信号接受和数据提取的效率和准确性，采用统计学方法加强现场数据收集的监测和质控，从而获得海量、多样且准确的结构化和非结构化数据。在数据处理阶段，设计高效、合理的数据库架构，对多源、异构数据的清洗、标准化和存储，最终形成满足数据分析需要数据，导出到统一的数据平台，是数据科学的独特优势。在数据分析阶段，既可采用传统的统计学方法描述数据的特征，也可通过数据挖掘和机器学习等大数据方法进行数据聚类分组、数据建模预测、文本挖掘等，深入发掘队列数据背后的规律。在研究结果展示和交流方面，通过可重复性研究定期生成数据报告，开发数据产品，对数据结果进行可视化呈现，开展基于数据驱动的决策等。

4. 大型队列研究的数据管理规范：随着大型队列研究在医学研究中的作用日益显现，研究者在加大人力和物力投入的同时，仍缺乏对队列研究的建设和利用的全面、系统的认识，尤其是在数据管理方面缺乏实践经验。在中华预防医学会的倡导下，北京大学联合中国医学科学院和北京理工大学组织专门的撰写小组，凝练总结了十余年队列建设的成功经验，完成了《大型人群队列研究数据处理技术规范》和《大型人群队列研究数据安全技术规范》[11-12]。这两个团体标准对不同来源、不同类型的队列数据标准化、清理、质控、整合、数据隐私性以及数据库安全管理进行规范化要求，适用于目前在国内建立的或拟开展大型人群队列研究，包括大型自然人群队列、区域性人群队列、针对某一特殊疾病或基于特殊机构开展的人群队列，也可供规模相对较小的人群队列研究参考。

四、应用前景与展望

近几十年来，我国居民生活习惯和环境发生了巨大的变化，人群健康和疾病谱也发生明显的改变，通过高质量、标准化的大型人群队列能准确地评估这些因素对健康的影响。近些年来，我国集中投入筹建了大量不同特征的人群队列，正逐步形成一个

具有遗传多样性、病例资源丰富的中国人群生物医学资源宝库。运用数据科学的方法和手段与队列研究开展过程中的各环节紧密结合，可以提高队列研究数据质量，优化数据清理、整合和开发，扩大和加深数据分析利用的广度和深度，生产出中国人群中的本土化证据，从而促进大型人群队列研究在揭示疾病的病因、发病机制以及提高早诊早治，开展循证卫生决策等方面的价值。

此外，大型人群队列研究的资源平台促进了研究数据和生物样本的科研合作与共享，跨学科优秀人才通力合作，不仅催生出一系列创新性的科研成果，也将培养一批具有国际视野的跨学科多层次科研人才队伍。

利益冲突　所有作者均声明不存在利益冲突

参 考 文 献

[1] Dhar V. Data science and prediction[J]. Commun ACM, 2013, 56(12): 64-73. DOI: 10.1145/2500499.

[2] Data Science Venn Diagram V2.0 [EB/OL]. (2014-01-06)[2018-10-20]. https://medium.com/3blades-blog/data-science-venn-diagram-v-2-0-c0250319e8f9.

[3] Davenport T H, Patil D J. Data scientist: the sexiest job of the 21st century[J]. Harv Bus Rev, 2012, 90(10): 70-76, 128.

[4] 唐金陵，Glasziou P. 循证医学基础[M]. 北京：北京大学医学出版社，2010.
Tang JL, Glasziou P. Essentials in evidence-based medicine[M]. Beijing: Peking University Medical Press, 2010: 45-58.

[5] 李立明，吕筠. 大型前瞻性人群队列研究进展[J]. 中华流行病学杂志，2015，36(11)：1187-1189. DOI：10.3760/cma.j.issn.0254-6450.2015.11.001.
Li LM, Lv J. Large prospective cohort studies: a review and update[J]. Chin J Epidemol, 2015, 36(11): 1187-1189. DOI: 10.3760/cma.j.issn.0254-6450.2015.11.001.

[6] 钱碧云，李森晶，张增利，等. 我国流行病学队列研究的现状与展望——2012年度预防医学学科发展战略研讨会综述[J]. 中国科学基金，2013，27(3)：138-142，157.
Qian BY, Li MJ, Zhang ZL, et al. Status and prospects of epidemiological cohort study in China — Summary of the 2012 strategic forum of preventive medicine[J]. Bull Natl Nat Sci Found Chin, 2013, 27(3): 138-142, 157.

[7] Framingham: the study and the town that changed the health of a generation [EB/OL]. (2018-10-10)[2018-10-20]. https://www.heart.org/en/news/2018/10/10/framingham-the-study-and-the-town-that-changed-the-health-of-a-generation.

[8] 徐萍，桂永浩，金力. 加快中国特色的精准医学的发展[J]. 中华儿科杂志，2016，54(5)：321-322. DOI：10.3760/cma.j.issn. 0578-1310.2016.05.001.
Xu P, Gui YH, Jin L. Speed up the development of precision medicine with Chinese characteristics[J]. Chin J Pediatr, 2016, 54(5): 321-322.

[9] 曲翌敏，江宇. 健康大数据的来源与应用[J]. 中华流行病学杂志，2015，36(10)：1181-1184.
Qu YM, Jiang Y. The sources and application of big data in healthcare[J]. Chin J Epidemiol, 2015, 36(10): 1181-1184. DOI: 10.3760/cma.j.issn.0578-1310.2016.05.001.

[10] The data science life cycle[EB/OL]. (2018-04-12)[2018-11-24]. https://datascience.berkeley.edu/about/what-is-data-science/.

[11] 中华预防医学会. 大型人群队列研究数据处理技术规范(T/CPMA 001-2008)[J]. 中华流行病学杂志，2019，40(1)：7-11. DOI：10.3760/cma.j.issn.0254- 6450.2019.01.003.
Chinese Preventive Medicine Association. Technical specification of data processing for large population-based cohort study (T/CPMA 001-2008)[J]. Chin J Epidemiol, 2019, 40(1): 7-11. DOI: 10.3760/cma.j.issn.0254-6450.2019.01.003.

[12] 中华预防医学会. 大型人群队列研究数据安全技术规范(T/CPMA 002-2008)[J]. 中华流行病学杂志，2019，40(1)：12-16. DOI：10.3760/cma.j.issn.0254-6450.2019.01.004.
Chinese Preventive Medicine Association. Technical specification of data security for large population-based cohort study (T/CPMA 002-2008)[J]. Chin J Epidemiol, 2019, 40(1): 12-16. DOI: 10.3760/cma.j.issn.0254- 6450.2019.01.004.

（收稿日期：2018-12-20）
（本文编辑：李银鸽）