标准·方案·指南

大型人群队列研究数据管理团体标准解读

余灿清[1] 刘亚宁[1] 吕筠[1] 卞铮[2] 谭云龙[2] 郭彧[2] 汤海京[3] 杨旭[3] 李立明[1]
[1]北京大学公共卫生学院流行病与卫生统计学系 100191；[2]中国医学科学院，北京 100730；[3]北京理工大学计算机学院 100081
通信作者：李立明，Email：lmleeph@vip.163.com

【摘要】 精准医学已成为我国科学技术优先发展的重点战略，大型人群队列的建设是人群疾病防治的重要资源，其研究结果为个体化治疗和精准预防提供科学证据。因此，如何规范化的建设大型人群队列是上述工作的基础。中华预防医学会组织北京大学等单位撰写《大型人群队列研究数据处理技术规范（T/CPMA 001–2018）》和《大型人群队列研究数据安全技术规范（T/CPMA 002–2018）》两项团体标准。标准以“科学性、规范性、可行性、可推广性”为原则，提出了大型人群队列研究在数据标准化技术、数据清理及质控技术、数据整合技术、数据隐私保护技术和数据库安全稳定性管理技术的原则和具体要求，以指导和规范我国已建立或拟开展的大型人群队列、区域性人群队列以及特殊人群队列，促进国内科研水平的提升，增加国际影响力，最大程度的支持疾病防控的决策与实践。

【关键词】 队列研究；数据处理；数据安全；技术规范；团体标准
基金项目：国家重点研发计划（2016YFC0900500，2016YFC0900504）
DOI：10.3760/cma.j.issn.0254–6450.2019.01.005

Interpretation for the group standards in data management for large population-based cohorts

Yu Canqing[1], Liu Yaning[1], Lyu Jun[1], Bian Zheng[2], Tan Yunlong[2], Guo Yu[2], Tang Haijing[3], Yang Xu[3], Li Liming[1]
[1]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing 100191, China; [2]Chinese Academy of Medical Sciences, Beijing 100730, China; [3]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081
Corresponding author: Li Liming, Email: lmleeph@vip.163.com

【Abstract】 Precision medicine became the key strategy in development priority of science and technology in China. The large population-based cohorts become valuable resources in preventing and treating major diseases in the population, which can contribute scientific evidence for personalized treatment and precise prevention. The fundamental question of the achievements above, therefore, is how to construct a large population-based cohort in a standardized way. The Chinese Preventive Medicine Association co-ordinated experienced researchers from Peking University and other well-known institutes to write up two group standards Technical specification of data processing for large population-based cohort study (T/CPMA 001–2018) and Technical specification of data security for large population-based cohort study (T/CPMA 002–2018), on data management. The standards are drafted with principles of emphasizing their scientific, normative, feasible, and generalizable nature. In these two standards, the key principles are proposed, and technical specifications are recommended in data standardization, cleansing, quality control, data integration, data privacy protection, and database security and stability management in large cohort studies. The standards aim to guide the large population-based cohorts that have been or intended to be established in China, including national cohorts, regional population cohorts, and special population cohorts, hence, to improve domestic scientific research level and the international influence, and to support decision-making and practice of disease prevention and control.

【Key words】 Cohort study; Data processing; Data security; Technical specification; Group standard
Fund programs: National Key Research and Development Program (2016YFC0900500, 2016 YFC0900504)
DOI：10.3760/cma.j.issn.0254–6450.2019.01.005

近些年来，随着生物医学研究的深入和发展，流行病学研究证据在疾病和健康问题中的研究价值日益体现。作为大型人群研究中证据价值最大、最可靠的研究资源，队列研究在流行病学病因学研究和疾病风险预测方面的价值也越来越受到人们的关注，国内外在人力、物力和财力等方面的投入越来越

大。然而，纵观我国队列研究的发展历史，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[1]。

为全面落实《国家中长期科学和技术发展规划纲要（2006—2020年）》的相关任务，“精准医学研究”于2016年被列为优先启动的重点专项之一，旨在以我国常见高发、危害重大的疾病及若干流行率相对较高的罕见病为切入点，实施精准医学研究的全创新链协同攻关，构建百万人以上的自然人群国家大型健康队列和重大疾病专病队列。

一、起草背景

大型人群队列研究的数据量大、来源复杂、类型丰富。研究内容上既包括问卷调查、体格检查、影像学检查和生化指标检测数据，也涵盖研究地区宏观地理和气候学信息、遗传学数据、终点事件数据等。因此，规范、有效地进行数据标准化和数据清理，完善多维、动态数据库的质量控制、数据整合及安全稳定是队列研究建设的重要课题。目前国内尚未形成与队列研究数据管理工作相关的国标、行标或专利性文件，缺少基于大队列这类多源、异构数据集的规范，更缺少对数据建设后的进一步规范。

2016年，国家重点研发计划“精准医学研究”资助重点专项“大型自然人群队列示范研究”项目（项目编号：2016YFC0900504）。该项目以中国慢性病前瞻性研究（China Kadoorie Biobank，CKB）项目[1-3]十余年的成功建设经验为基础，建立大型人群队列研究的规范化操作流程，制定人群队列的建设标准。此次撰写并发布的两项团体标准《大型人群队列研究数据处理规范（T/CPMA 001–2018）》[5]与《大型人群队列研究数据安全规范（T/CPMA 002–2018）》[6]（标准）就是对大型人群队列研究的数据管理方面提出了标准和要求。

二、前期工作基础

CKB队列是我国迄今为止规模最大且全球领先的大型人群队列，项目建设引入了国际先进的管理理念和技术手段，坚持标准化操作规范，实行全程计算机化管理。从现场数据采集到血样的登记、分装、储存、运输以及材料的供应和运输、交流通信、与死亡和发病数据关联等各个环节，项目通过专用软件系统进行规范化管理，为提高工作效率、同步和动态化的现场质量控制创造基本条件。由此可见，项目的管理模式和理念均处于国际先进水平，而且随着项目进行不同阶段，可以预见并解决大型队列研究的问题，这些实践经验可指导其他人群队列的建设，提高研究质量和工作效率，创新项目管理运行机制。

在此基础上，CKB项目团队以国家科技支撑计划项目《区域人口健康大型队列适宜技术的标准化与应用》为依托，着眼于大数据时代及超大型队列研究不断发展的趋势，对大型队列基线调查中的各类信息和样本收集的规范化方式及流程进行了系统的梳理总结，出版了《大型人群队列研究调查适宜技术》[7]和《大型人群队列研究随访监测适宜技术》[8]两本专著，重点介绍了基于我国国情的大型人群队列研究中主要慢性病发病、死亡监测的适宜技术及相应的技术规程。两本书稿的出版，对我国大型队列研究整体设计和实施各个环节中的标准进行了规范化要求，为国内不同水平的队列研究建设提供了示范和技术支撑，也为国家层面上的标准统一、数据规范和资源共享奠定了基础。

三、编写原则

本标准为首次制定，遵循“科学性、规范性、可行性、可推广性”的原则，根据国家数据标准化及信息安全相关法律、法规和法规性文件，行业标准及规范的要求，结合大型人群队列数据库管理的规范化研究成果，以及现有队列建设的实践经验进行标准编制。

本标准旨在制定符合国情、满足目标人群和地域特点、可操作性强、可推广的人群队列建设的团体标准和规范化操作流程，指导其他人群队列的建设，促进国内科研水平的提升，增加国际影响力，最大程度的支持疾病防控的决策与实践。

四、团标内容

本次制定的两项标准规定了大型人群队列研究在数据管理各个环节的基本原则和技术规范，适用于各级卫生行政部门、各级各类医疗卫生机构以及科研机构等规范开展大型人群队列研究过程中与数据管理有关的工作。

两个标准的主要章节相似，包括范围、规范性引用文件、术语和定义、技术规范要求以及相关参考文献，其中技术规范的主要内容包括六个方面：

1. 数据标准化技术：结合数据利用要求，将数据按照相关行业的要求进行标准化，提出数据标准化的基本要求，介绍数据标准化的实施过程、数据库设计要求以及不同类型数据的标准化方法等一系列操作规范。

2. 数据清理及质控技术：结合大型自然人群队列的多来源数据（包括现场调查数据、长期随访监测数据、组学数据等），从数据检查、问题处置和统计学监测三个方面制定有针对性的数据清理标准，并对数据质量控制提出要求。

3. 数据整合技术：根据数据的特点，规定了结构化或非结构化数据整合的基本过程，将多源异构数据整合到标准数据库中，便于数据挖掘及分析。

4. 数据处理记录与报告：数据管理要求对上述三个阶段所涉及的数据处过程、依据和结果进行详细的记录，并定期报告和存档。

5. 数据的隐私保护：根据伦理学的要求和现场工作的实践，通过加密和其他安全措施，保护受试者的基本利益。本标准规定了大型队列研究数据隐私保护的基本要求，提出了数据隐私的类型、数据隐私参与角色及主要环节的隐私保护措施。

6. 数据安全和稳定性管理：大型人群的队列研究数据量较大，数据的安全和稳定是项目实施的基本保障，这部分内容从基本技术要求、操作规范、存储和备份等环节提出建设和管理要求，确保数据的安全和稳定。

五、团标使用的注意事项

1. 结合研究目的，进行数据管理的顶层设计：本标准的制定遵循“可推广性”原则，考虑大型人群队列数据管理的一般情况，规定了大型人群队列数据管理的主要原则和规范化操作要求。因此，研究者在使用本标准时，需结合队列自身的研究目的和实际需求，科学合理地制定队列研究的数据管理细则。

需要强调的是，制定这些细则需要先于队列建设，建立好数据管理有关的规范化操作流程，有利于快速、高效、准确的收集队列数据。对于正在开展的队列研究，建议对照着本标准的内容，逐条比对，查漏补缺，确保数据管理过程的连贯性和规范性。

2. 理解标准要求，因地制宜地选择数据管理技术：本标准介绍了大型人群队列数据管理过程中多种常用的技术手段，但并非同时适用于所有的人群队列。项目应结合队列自身的要求和实际情况，充分理解各种技术手段的原理、实施方法和优缺点，深刻理解标准要求，从准确性、可及性、可行性、经济性等方面综合考虑，选择最适宜的一种或多种技术手段和相适应的数据方式，以达到队列数据的最优化管理。

3. 及时更新数据管理技术，与时俱进地开展队列建设：大型人群队列研究数据量大、来源复杂、类型丰富，通常涉及多领域的研究内容。因此，需要开阔视野，综合利用多种技术交叉融合。近些年来，数据科学发展势头迅猛，可以采用多种计算机科学技术、数理统计方法等途径完成数据管理和质控工作，充分发挥本标准的作用和优势。同时，还需要密切关注国家制定的数据标准化及信息安全相关法律、法规和法规性文件，行业标准及规范等，及时遵守和实施文件的具体要求。

六、展望

目前，我国正处于快速、批量建设大规模人群队列研究时期，不同团队对于队列数据管理的理解迥异。本标准旨在制定符合国情、满足目标人群和地域特点、可操作性强、可推广的人群队列建设的行业标准和规范化操作流程，指导这些人群队列的建设。

本标准的提出和应用，不仅有利于统一、规范地建立一批高质量的队列资源，生产高质量的本土化证据，最大程度的支持疾病防控的决策与实践，而且为今后不同团队的队列数据互联互通打下基础，资源共享，促进高水平研究成果的发表，提升我国学者在国际学术舞台上的地位和影响力。

利益冲突 所有作者均声明不存在利益冲突

参 考 文 献

[1] 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.

[2] 李立明，吕筠，郭彧，等. 中国慢性病前瞻性研究：研究方法和调查对象的基线特征［J］. 中华流行病学杂志，2012，33(3)：249-255. DOI：10.3760/cma.j.issn.0254-6450.2012.03.001.
Li LM, Lv J, Guo Y, et al. The China Kadoorie Biobank: related methodology and baseline characteristics of the participants［J］. Chin J Epidemiol, 2012, 33(3): 249-255. DOI: 10.3760/cma.j.issn.0254-6450.2012.03.001.

[3] Chen ZM, Chen JS, Collins R, et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up［J］. Int J Epidemiol, 2011, 40(6): 1652-1666. DOI: 10.1093/ije/dyr120.

[4] Chen ZM, Lee L, Chen JS, et al. Cohort profile: The Kadoorie study of chronic disease in China (KSCDC)［J］. Int J Epidemiol, 2005, 34(6): 1243-1249. DOI: 10.1093/ije/dyi1d74.

[5] 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 001-2018)［J］. 中华流行病学杂志. 2019, 40(1): 7-11. DOI：10.3760/cma.j.issn.0254-6450.2019.01.002.
Chinese Preventive Medicine Association. Technical specification of data processing for large population-based cohort study (T/CPMA 001-2018)［J］. Chin J Epidemiol, 2019, 40(1): 7-11. DOI: 10.3760/cma.j.issn. 0254-6450.2019.01.002.

[6] 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 002-2018)［J］. 中华流行病学杂志. 2019, 40(1): 12-16. DOI：10.3760/cma.j.issn.0254-6450.2019.01.003.
Chinese Preventive Medicine Association. Technical specification of data security for large population-based cohort study (T/CPMA 002-2018)［J］. Chin J Epidemiol, 2019, 40(1): 12-16. DOI: 10.3760/cma.j.issn.0254-6450.2019.01.003.

[7] 李立明. 大型人群队列研究调查适宜技术［M］. 北京：人民卫生出版社，2014.
Li LM. Suitable instigation techniques for large population-based cohort study［M］. Beijing: People's Medical Publishing House, 2014.

[8] 李立明. 大型人群队列随访监测适宜技术［M］. 北京：中国协和医科大学出版社，2015.
Li LM. Suitable techniques of follow-up and surveillance for large population-based cohort study［M］. Beijing: Peking Union Medical College Press, 2015.

（收稿日期：2018-12-13）
（本文编辑：李银鸽）