·中国国家出生队列·

云端信息平台在中国国家出生队列建设与研究中的应用

杜江波[1]　陶诗瑶[1]　林苑[2,3]　赵杨[4]　吕红[1]　夏彦恺[2,5]　陆春城[2,5]　吴炜[2,5]
马红霞[1,2]　靳光付[1,2]　胡志斌[1,2]　沈洪兵[1,2]
[1]南京医科大学公共卫生学院流行病学系 211166；[2]南京医科大学生殖医学国家重点实验室 全球健康研究中心 211166；[3]南京医科大学公共卫生学院儿少卫生与妇幼保健学系 211166；[4]南京医科大学公共卫生学院生物统计学系 211166；[5]南京医科大学公共卫生学院教育部现代毒理学重点实验室 211166
通信作者：沈洪兵，Email：hbshen@njmu.edu.cn

【摘要】 大型出生队列是持续、动态地收集个体生命早期暴露信息，探讨暴露与生命远期健康结局因果关联的一种重要的队列研究类型。但由于其设计复杂、实施难度大，如何保证出生队列建设的高质量高效率是国内外流行病学研究者面临的主要挑战。2016年，国家重点研发计划资助的中国国家出生队列（China National Birth Cohort）建设正式启动。该队列在设计实施过程中，不断积累经验，充分运用网络和信息化手段，探索并建立了一套“云端信息平台”，以支撑覆盖全国16家单位的出生队列建设。经过四年的发展，该系统平台在“出生队列人群招募和随访管理、数据的实时交互、队列质量控制、多级权限管理和职能划分”等多个方面已经发展出一整套完善的建设方案。该平台的设计框架和功能要素对于我国今后出生队列乃至大型人群研究的信息化建设具有重要的参考意义。

【关键词】 出生队列；　信息化平台；　质量控制
基金项目：国家重点研发计划（2016YFC1000200）

Application of cloud-based information platform in China National Birth Cohort

Du Jiangbo[1], Tao Shiyao[1], Lin Yuan[2,3], Zhao Yang[4], Lyu Hong[1], Xia Yankai[2,5], Lu Chuncheng[2,5], Wu Wei[2,5], Ma Hongxia[1,2], Jin Guangfu[1,2], Hu Zhibin[1,2], Shen Hongbing[1,2]
[1]Department of Epidemiology, School of Public Health, Nanjing Medical University, Nanjing 211166, China; [2]State Key Laboratory of Reproductive Medicine, Center for Global Health, Nanjing Medical University, Nanjing 211166, China; [3]Department of Maternal, Child and Adolescent Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China; [4]Department of Biostatistics, School of Public Health, Nanjing Medical University, Nanjing 211166, China; [5]Key Laboratory of Modern Toxicology of Ministry of Education, School of Public Health, Nanjing Medical University, Nanjing 211166, China
Corresponding author: Shen Hongbing, Email:hbshen@njmu.edu.cn

【Abstract】 Birth cohort is an important observational study which can continuously and dynamically collect the exposure changes and health outcomes from gametophyte development to adolescence and even old age. However, because of its complex design and difficult implementation, how to construct birth cohort with high quality and high efficiency is the main difficulty faced by epidemiologists at home and abroad. In 2016, China National Birth Cohort was officially launched. The network and information technology were used to explore, and a set of "cloud-based information platform" was established to support this queue construction, containing 16 units in

DOI：10.3760/cma.j.cn112338-20201211-01404
收稿日期　2020-12-11　　**本文编辑**　李银鸽
引用本文：杜江波，陶诗瑶，林苑，等. 云端信息平台在国家出生队列建设与研究中的应用[J]. 中华流行病学杂志，2021，42(4)：586-590. DOI：10.3760/cma.j.cn112338-20201211-01404.

China. After four years of development, the platform has formed a complete set of programs about the construction of cohort information platform, which including recruitment and follow-up management of participants, real-time data interaction, queue quality control, multi-level authority management and function division. The relevant design framework and functional elements provide the references to the future information construction of large-scale birth cohort and even population-based research in China.

【**Key words**】 Birth cohort;　Information platform;　Quality control

Fund program: National Key Research and Development Program of China (2016YFC1000200)

近年来，信息化建设对我国社会诸多行业和领域的高质量发展均发挥了重要的促进作用[1-2]。队列研究是以前瞻性采集和分析大样本人群数据为核心内容的重要流行病学研究方法，在数据获取、清理和分析等多个环节都存在丰富的网络信息技术应用场景。因此，应用信息化的手段提升队列建设质量和效率正在成为近年来大型队列建设领域的重要发展趋势[3]。

出生队列是一种特殊类型的队列研究设计，旨在探讨环境行为因素等生命早期暴露对亲代生殖生育结局和子代近、远期健康的潜在影响[4]。近年来，胚胎发育过程所经历的有害因素可能对子代远期健康状况具有不利影响，即“疾病胚胎起源学说”正被越来越多研究者所接受[5]。因此，国内外出生队列研究得以快速发展[6-9]。2016 年，由国家重点研发计划资助的国家出生队列项目正式启动。该项目的目标是建成全国多中心的大样本前瞻性出生队列，并基于队列开展流行病学研究。与此同时，该队列建设在多中心协同、数据高质量即时汇交、队列数据资源开放共享途径等方面都面临前所未有的挑战。此外，出生队列特有的孕期多时点随访、多种类数据和生物样本采集、以家庭为随访单位等特点，也为高质量的队列随访带来了很大的困难。

针对以上大型出生队列实施过程中面临的难题，本项目实施团队通过反复调研和长期实践，独立自主地研发了一套基于云端的出生队列信息化平台。经过近四年的反复优化和迭代升级，该信息化平台在队列成员管理、信息采集、跟踪随访、质量控制等多个方面具备了鲜明特色和独特优势。主要包括队列成员管理系统、问卷编辑系统和共享合作平台，其中队列成员管理系统设置了不同的模块以供调查员采集调查对象基线及随访信息，也可供后台管理者导出问卷并质控；问卷编辑系统主要供管理者添加问卷及设置问题逻辑；共享合作平台则专门用于储存定库数据，以便申请者随时调用分析。本文将系统介绍云端信息化平台的主要设计构架和功能特点（图 1），以期为国内外其他大型复杂队列的建设提供参考。

一、智能化的人群招募和随访管理

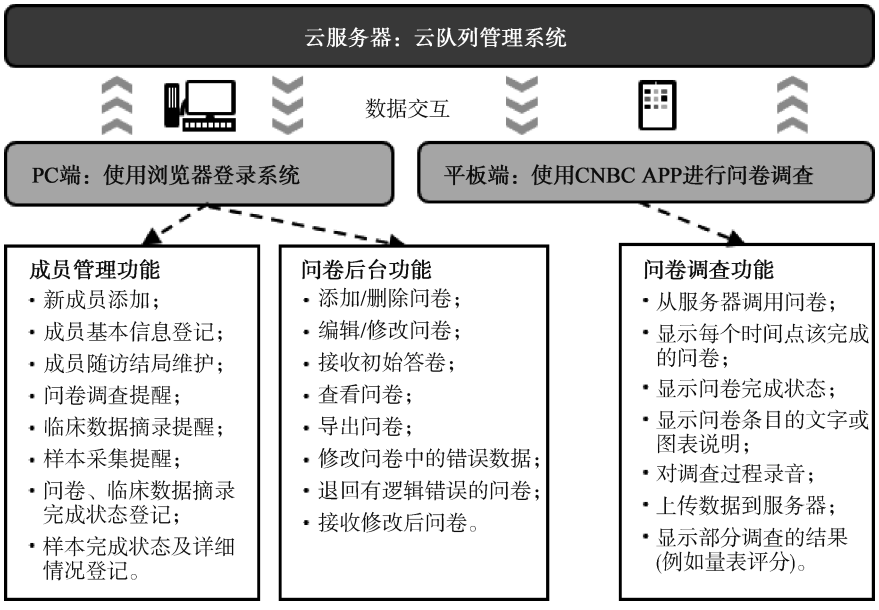

图 1　云队列管理系统的组成和功能

中国国家出生队列涵盖了12个省（自治区、直辖市）的16家单位，从母亲孕前到子代3岁期间设置了10余个随访节点。因此，多中心的协同和不同随访时点同步是国家出生队列实际工作中首要解决的问题[10]，因此，基于信息化平台的建设方案包括：①各中心用户可通过统一的平台界面在线编辑队列成员基本信息表：该功能使得项目组能够实时地掌握各分中心的研究对象招募和随访的进程。②问卷智能化编辑和推送功能：该功能允许项目组统一配置调查问卷的内容、指定调查问卷需要完成的时间区间，并根据内置逻辑在平板电脑等终端实现不同随访节点问卷智能化推送。调查对象只需登录其队列成员ID即可自动获取该时点需要完成的问卷，问卷完成后数据实时汇交至总项目组。除此之外，如某份问卷需要更新，主中心可通过该模块及时将新版问卷推送至各个调查终端，从而保证了多中心同步。③智能化工作提醒：队列中调查员往往需要提前与纳入对象联系确定随访时间或随访结局等，系统可以通过预设逻辑，自动拟定调查员所需维护的队列成员信息和所需开展的随访工作清单，避免了人工安排随访日期和内容，极大地提高了队列随访工作的质量和效率。

上述功能模块的应用使得出生队列的招募和随访工作摆脱了地域限制，完全实现了队列招募和随访工作同步化，并且随访工作的具体时间表均由计算机自动生成，帮助工作人员更有规划地安排每日随访工作，避免遗漏。

二、实时化的电子问卷调查和数据质量控制

传统的流行病学研究主要通过纸质调查问卷采集人群数据[11]。然而，这种方法已经难以适应大型复杂人群队列的建设。中国国家出生队列涉及6万个家庭成员多个时间节点的随访调查，累计需汇集数以百万计的问卷调查和临床检查数据，如使用传统纸质调查不仅耗时费力，且质量难以控制，可行性极低。国外有研究者分析发现，与纸质调查法相比，电子化调查可节省49%~62%的经济成本，并且电子化调查在数据管理阶段明显优于纸质调查[12]。基于信息化平台的建设方案包括：①问卷数据实时汇交：调查对象改用平板电脑等采集终端完成问卷调查，并且调查问卷的结果数据、调查问卷的“.txt”原始文件以及问卷完成时的录音文件均可实时上传至项目组的云端服务器，一方面节省了大量的纸质调查表，大大缩短了“数据采集-数据清理-数据分析”这一核心过程的时间，另一方面也对调查过程的原始数据文档做了及时备份；②及时、高效的数据质量控制：借助信息化平台，项目组可以对全国多中心汇交的数据及时开展数据质量控制，对于错误、缺失、存疑的调查问卷，可通过平台将其及时退回至分中心，请其说明缘由并要求其核对，分中心补充核实后需再次提交至项目组。通过数据质量审核的调查问卷方可归档进入分析环节。上述流程最终形成了闭环的工作模式，最大程度保障了数据的准确性。

三、多维度的队列质量控制

大型出生队列的数据具有大样本量、数据维度多、随访节点密集等特点[13]。任何关乎数据质量的微小错误都可能严重影响后续结果结论的真实性。因此，严格的质量控制应贯穿队列信息获取和整理的各个环节。基于信息化平台，国家出生队列数据质量控制从多个维度开展。

1. 问卷设计环节：针对每一道题目，对其输入的文本格式进行限定，也可设置警示值范围和异常值范围，对于必填问题可设置必填限定，这些操作可以最大限度地保证从数据入口端提升数据采集的完整性和准确性。

2. 数据采集环节：数据采集过程的规范性是传统纸质调查问卷中最难核查的部分，但基于信息化平台则可以多角度核查问卷调查的过程质量。例如，后台质控专员可以计算问卷完成时长，找出极端值，对其数据质量进行核查，也可以通过抽查重听问卷录音文件有效地监督调查员在调查过程中的提问内容和方式。基于系统的质控可以在获得实时上传的调查数据后即刻进行质控分析，针对发现的问题也能够在第一时间将异常数据退回给相应的调查员，从而在最短的时间内对异常数据进行核实和更正。

3. 数据使用环节：总中心质控专员对导出问卷数据库进一步逻辑核查，如对于量表中所有问卷的答案选项均相同、出生日期小于妊娠日期、每日睡眠时间超过16 h等细节。这类数据逻辑问题通常情况下难以发现，而基于信息化平台的内置程序代码可轻松高效地完成问卷的深度逻辑核查，大大提升了最终进入分析环节的数据可靠性。

4. 生物样本质控：基于信息化平台可以对所采集生物样本的体积、颜色、冷链转运执行情况等与样本质量密切相关的重要参数进行质量控制。这些质控对于后期开展高质量的生物样本检测具有重要参考价值。

四、精细化的权限管理和职能划分

中国国家出生队列分中心遍布全国，各分中心均已配备1~2名管理员和若干名调查员、系统维护员、样本处理员等。根据地域分布这些中心被划分至不同课题组，各课题组长单位的管理员负责统筹本组负责区域内各分中心队列建设任务和进展。项目组长通过与各课题组长单位的联络协调机制，全面掌握全国多中心队列建设任务的推进情况。此外，各分中心队列纳入现场和纳入人群类型多样，纳入现场包括医院产科、生殖中心以及社区医院，纳入人群类型包括辅助生殖和自然妊娠家庭。总之，中国国家出生队列建设工作的参与者众多，分工和协调机制复杂。针对这一现象，队列信息化平台采用精细化的用户权限分级，实行项目管理员、课题组管理员、分中心管理员、分中心操作员的四级权限管理，严格限制操作员的数据导出权限，即仅能导出工作相关任务清单，不能导出队列数据库，管理员具有查看和导出本项目组、本课题组或本中心的队列数据的权限，从而确保了国家出生队列建设的大团队在精细分工、高度协同的机制下开展工作。

五、便捷高效、安全可控的数据管理

协同共建、开放共享、合作共赢是中国国家出生队列始终秉承的核心理念。然而，目前我国大型队列数据开放共享的机制和模式尚不成熟，发展相对落后[14-17]。因此，中国国家出生队列在队列建设的同时，也将探索大型队列数据资源开放、共享、合作的相关机制和模式。目前，中国国家出生队列项目组从原则、流程、技术等多个方面进行了完善的顶层设计，并且在实践过程中不断优化总结，已经形成了较为成熟的方案。研究者可通过中国国家出生队列网站在线提交课题申报书和数据分析权限申请。对于符合创新性和可行性的申请，队列平台将通过在线授权和远程访问的方式为申请人提供数据分析权限，实现了数据共享途径的便利化和高效化。

数据安全对于队列研究项目至关重要。特别是基于信息化平台的队列研究项目，数据安全保障措施应该更加严格。国家出生队列制定了全面稳妥的数据安全保障措施：①严格执行数据脱敏，即用于识别队列成员的身份信息和联系方式等隐私数据与队列研究数据完全独立保存管理，两类数据库连接所需对应“编码钥匙”由专人保存；②为队列云端服务器配置全面的防火墙等多重防攻击技术屏障；③用户访问服务器采用https协议，即通过传输加密和身份认证保证数据传输过程的安全性；④队列所有电子数据均按照加密存储、多重拷贝、异地备份的方式避免由存储介质丢失、损害等造成的数据安全风险；⑤对数据接触者定期开展数据安全培训和考核，避免人为原因造成的数据泄露。

六、总结与展望

“大数据”时代已经到来，传统流行病学研究必须寻求与大数据、互联网等先进信息化技术和手段的结合。无纸化的云端信息平台将会越来越多地应用于大型队列研究。中国国家出生队列的信息系统在我国队列建设领域的信息化方面做了率先尝试，因此也在系统的设计构架、功能特点等方面体现出一些特点、优势和标准体系[18]。该系统在问卷数据的实时交互等方面与国际上其他队列的经典做法相比有一定特色，但在数据采集过程的质控、质量保证和IT团队支持等方面还需要不断学习国外先进做法[19]。本文结合中国国家出生队列建设与研究的实践经验，从智能化的出生队列人群招募和随访管理、实时化的电子问卷调查和数据质量控制交互、多维度的队列质量控制、精细化的权限管理和职能划分、便捷的队列资源共享与合作五个方面，对国家出生队列云端信息化平台建设的经验进行阐述，希望为其他出生队列或大型人群研究项目的信息化建设方案提供有价值的参考。

利益冲突　所有作者均声明不存在利益冲突

参　考　文　献

[1] 湛永乐，岳和欣，石英杰，等. 基于妇幼保健网络建立母婴队列的可行性分析[J]. 中华流行病学杂志，2020，41(4)：605-610. DOI:10.3760/cma.j.cn112338-201900726-00553.
Zhan YL, Yue HX, Shi YJ, et al. Feasibility on the development of maternal and child cohorts, based on the maternal and child care network[J]. Chin J Epidemiol, 2020, 41(4): 605-610. DOI: 10.3760/cma. j. cn112338-201900726-00553.

[2] 叶荣伟，廖传颉，李松，等. 生育健康监测的电子化研究[J]. 中华流行病学杂志，2001，22(3):166-168.
Ye RW, Liao CJ, Li S, et al. The establishment of an electronic reproductive health surveillance system (ERHSS) [J]. Chin J Epidemiol, 2001, 22(3):166-168.

[3] 刘勇，王丽敏，彭永祥，等. 多中心血糖检测电子化质量监控系统的建立与实施[J]. 中华流行病学杂志，2015,36(5)：506-509. DOI:10.3760/cma.j.issn.0254-6450.2015.05.020.
Liu Y, Wang LM, Peng YX, et al. Designing and implementation of a web-based quality monitoring system for plasma glucose measurement in multicenter

population study[J]. Chin J Epidemiol, 2015,36(5):506-509. DOI:10.3760/cma.j.issn.0254-6450.2015.05.020.

[4] 王磊, 孙蕾, 何晓燕, 等. 中国出生队列研究进展[J]. 中华流行病学杂志, 2017, 38(4): 556-560. DOI: 10.3760/cma. j. issn.0254-6450.2017.04.029.
Wang L, Sun L, He XY, et al. Birth cohort studies in China:a review[J]. Chin J Epidemiol, 2017, 38(4): 556-560. DOI: 10.3760/cma.j.issn.0254-6450.2017.04.029.

[5] Hanson M, Gluckman P. Developmental origins of noncommunicable disease: population and public health implications [J]. Am J Clin Nutr, 2011, 94, 6 Suppl: S1754-1758. DOI:10.3945/ajcn.110.001206.

[6] Golding J, ALSPAC Study Team. The Avon Longitudinal Study of Parents and Children (ALSPAC)-study design and collaborative opportunities [J]. Eur J Endocrinol, 2004, 151 Suppl 3:U119-123. DOI:10.1530/eje.0.151u119.

[7] Jaddoe VW, van Duijn CM, Franco OH, et al. The Generation R Study: design and cohort update 2012 [J]. Eur J Epidemiol, 2012, 27(9): 739-756. DOI: 10.1007/s10654-012-9735-1.

[8] Qiu X, Lu JH, He JR, et al. The Born in Guangzhou Cohort Study (BIGCS) [J]. Eur J Epidemiol, 2017, 32(4):337-346. DOI:10.1007/s10654-017-0239-x.

[9] Huang JV, Leung GM, Schooling CM. The association of air pollution with birthweight and gestational age: evidence from Hong Kong's 'Children of 1997' birth cohort [J]. 2017, 39(3):476-484. DOI:10.1093/pubmed/fdw068.

[10] 吴美琴, 吴宇航, 赵丽, 等. 多中心间的协同性决定队列生物样本的一致性 [J]. 中国医药生物技术, 2015, 10(6): 489-493. DOI:10.3969/j.issn.1673-713X.2015.06.003.
Wu MQ, Wu YH, Zhao L, et al. The coordination between multiple centers determines the consistency of biological samples in the queue[J]. Chin Med Biotechnol, 2015, 10 (6):489-493. DOI:10.3969/j.issn.1673-713X.2015.06.003.

[11] 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.0254-6450.2019.01.004.
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. 2019.01.004.

[12] Pavlović I, Kern T, Miklavcic D. Comparison of paper-based and electronic data collection process in clinical trials: Costs simulation study [J]. Contemp Clin Trials, 2009, 30(4):300-316. DOI:10.1016/j.cct.2009.03.008.

[13] 余灿清, 李立明. 大型队列研究中的数据科学[J]. 中华流行病学杂志, 2019, 40(1):1-4. DOI:10.3760/cma.j.issn.0254-6450.2019.01.001.
Yu CQ, Li LM. Data science in large cohort studies[J]. Chin J Epidemiol, 2019, 40(1): 1-4. DOI: 10.3760/cma. j. issn. 0254-6450.2019.01.001.

[14] Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials [J]. BMJ, 2017, 357: j2372. DOI:10.1136/bmj.j2372.

[15] 董文斌, 雷小平. 大数据时代出生队列研究的新趋势 [J]. 西部医学, 2015, 27(5): 641-644. DOI: 10.3969/j. issn. 1672-3511.2015.05.001.
Dong WB, Lei XP. The new trends of birth cohort study in big data era [J]. Med J West China, 2015, 27(5): 641-644. DOI:10.3969/j.issn.1672-3511.2015.05.001.

[16] 胡志斌. 建设高质量出生队列支撑全生命周期人群健康研究 [J]. 中华预防医学杂志, 2018, 52(10): 973-975. DOI: 10.3760/cma.j.issn.0253-9624.2018.10.002.
Hu ZB. Constructing high-quality birth cohort and supporting the study on full life cycle population health [J]. Chin J Prev Med, 2018, 52(10):973-975.DOI:10.3760/cma.j.issn.0253-9624.2018.10.002.

[17] 王慧, 张作文. 我国人群队列研究的现状、机遇与挑战[J]. 中华预防医学杂志, 2014, 48(11): 1016-1021. DOI: 10.3760/cma.j.issn.0253-9624.2014.11.01.
Wang H, Zhang ZW. Current situation, opportunities and challenges of population cohort research in China[J]. Chin J Prev Med, 2014, 48(11):1016-1021.DOI:10.3760/cma.j.issn.0253-9624.2014.11.01.

[18] 中华预防医学会. TCPMA 015.3–2020 出生队列技术规范 成员信息系统[EB/OL]. (2020-05-01) [2020-12-01]. https://max. book118. com/html/2020/1231/7040012030003036.shtm.
Chinese Preventive Medicine Association. TCPMA 015.3-2020 Birth cohort technical specification. Part 3: participants information system[EB/OL]. (2020-05-01) [2020-12-01]. https://max. book118. com/html/2020/1231/7040012030003036.shtm.

[19] Chen ZM. Population Biobank Studies: A Practical Guide [M]. Springer, 2020: 145-169.