·学术动态·

大型前瞻性队列研究实施现况及其特点

熊玮仪 吕筠 郭彧 李立明

【关键词】 大型前瞻性队列；流行病学

Overview on the practice and characteristics of large prospective cohort studies

Xiong Weiyi[1], *Lyu Jun*[2], *Guo Yu*[3], *Li Liming*[2,3]. 1 *Division of Epidemiology, School of Public Health and Primary Care, the Chinese University of Hong Kong, Hong Kong, China*; 2 *Department of Epidemiology and Biostatistics, School of Public Health, Peking University*; 3 *Chinese Academy of Medical Sciences*

Corresponding author: *Li Liming*, Email: lmlee@pumc.edu.cn

This work was supported by a grant from the National Science and Technology Support Projects for the "Twelve Five-Year Plan" of China (No. 2011BAI09B01).

【Key words】 Large prospective cohort; Epidemiology

大型前瞻性队列(大型队列)是20世纪中叶以来随着对慢性病发病机制研究的不断深入，而逐渐兴起的现代病因研究项目。在流行病学研究设计的分类体系中，大型队列仍属于队列研究，但绝非普通规模队列的简单放大[1]。

一、诞生背景

随着相关研究的深入，人们发现大多数疾病，尤其是慢性疾病(肿瘤、心血管疾病等)的发病机制复杂，不单纯由某一先天遗传特征或后天环境暴露引起，而是由基因、环境、行为等多因素联合作用导致[2]。此类疾病已成为现代社会最主要的疾病负担，并造成巨大的健康和经济损失[3,4]。深入揭示上述因素在慢性病发生、发展中的单独和/或交互作用，以寻求适当的干预环节，已成为现代病因研究的重要课题。

流行病学有多种病因研究设计，但并非所有都适用于上述疾病。当疾病的潜隐期(latency period)较长、出现明显临床症状的进展较慢(数年甚至数十年)时，若开展病例对照研究，则回忆偏倚对研究结果的干扰较大且难以排除；当疾病在人群中的发生率极低时，则普通规模的队列研究在相当长时间内均无法累积足够数量病例，尤其是在研究基因-环境交互作用时[5-9]。故对于此类疾病，只有用足够大的样本、随访足够长的时间，才能捕捉到足量病例；且只有在疾病发生之前就采集到准确的暴露信息，并及时、全面掌握研究人群中真实的结局信息，方能让基本无偏的基因、环境、基因-环境交互研究成为现实[7]。此为大型队列的设计初衷。

二、规模要求

多大规模称之为"大型队列"，目前尚无定论。就研究设计而言，具体研究的规模大小，应根据研究目的、内容而定，一些学术组织和机构也提供了推荐的样本量，例如对于建立生物样本库(biobank)，P³G(Public Population Project in Genomics and Society)网站对"大规模人群生物样本库(large population-based biobank)"的最低样本量要求为1万名健康人群[10]；对于研究遗传、环境因素对于慢性病的弱效应及其交互作用，美国国家人类基因组研究所(National Human Genome Research Institute)的推荐样本量为50万[7]。目前公认的大型队列研究样本量多在5万至50万之间，其中以10万以上居多。

三、实施现况

建立和维护大型队列需要有巨大的投入，对于研究地区的医疗卫生服务基础、信息技术(IT)应用程度以及研究机构的生物样本处理、基因测序和数据分析能力等也有较高要求。因此，现有的大型队列多建立在社会经济发展总体水平较高的国家或地区。目前在P³G网站注册的研究共有164项，大多位于北美和欧洲，我国大陆地区有8项[10]。

1. 模式概述[1,9]：大型队列按其来源可分为两类。一类为在已有较小规模(相对大型队列而言)的研究基础上，采取合并、协作、追加样本量等方式建立的队列，如欧洲的European Prospective Investigation into Cancer and Nutrition (EPIC)；另一类是新建的队列，如英国的UK Biobank、美国的Nurses' Health Study (NHS)、中国的China Kadoorie Biobank (CKB；或Kadoorie Study of Chronic Disease in China)。具体操作模式分为三种，即分散模式(decentralized)、集中模式(centralized)和单一中心(single site)[9]。建立在已有研究基础上的大型队列多为分散模式，新建的大型队列多采用后两种模式。

(1)分散模式：即传统的多中心联合/协作研究(multisite consortia)。研究设有一个协调中心(coordinating center)以及若干分中心，根据分中心所在位置划分为若干地理区域。协调中心负责总体设计、质控、培训、数据汇总及核查等；各区域内的研究，如研究对象招募、基线调查、随访、重复调查等，均交由分中心独立完成，包括采样及检测。所需人员和仪器多由分中心自行筹措和管理，所得的数据和样本也多由分中心各自保存。此模式的实施前提为符合资质的机构(分中心)愿意加入本项研究。其优点是可以充分发挥分中心的作用，以及建立协调中心的投入相对较小；缺点为难以保证

DOI: 10.3760/cma.j.issn.0254-6450.2014.01.022

基金项目：国家"十二五"科技支撑计划(2011BAI09B01)

作者单位：香港中文大学公共卫生及基层医疗学院流行病学系(熊玮仪)；北京大学公共卫生学院流行病与卫生统计学系(李立明、吕筠)；中国医学科学院(李立明、郭彧)

通信作者：李立明，Email: lmlee@pumc.edu.cn

数据的标准化和可比性，总费用高昂。此种模式的范例为EPIC研究。

(2)集中模式：集中模式的大型队列也设有一个协调中心和若干分中心，但职责划分与分散模式有很大不同。集中模式的分中心是实地开展问卷调查、体格检查和采样的场所，虽然有可能建立于某机构内，但对于研究而言只是临时设置，即根据研究需要，临时指定，由协调中心提供设备、派出或在当地雇佣人员，任务完成后本项目的人员和设备即撤除。该模式的特点是分中心不储备数据和样本，所有调查数据实时或在完成调查后短期内即传送至协调中心，样本采集后也及时转运至协调中心或指定实验室。协调中心在集中模式中发挥了至关重要的作用，负责设计、执行和全程管理，同时也是数据和样本的保管中心。虽然该模式建立和维护协调中心的费用较高，但由于分中心的运行成本低，因此节约了总成本。集中模式的范例为UK Biobank研究，我国CKB研究也借鉴了此模式[11]。

(3)单一中心：由一个机构负责所有的招募、基线调查、随访、重复调查以及采样检测等。以此种方式建立的大型队列，势必依托于某一成熟、强大的研究机构或医疗集团，如美国 Kaiser Permanente Research Program on Genes, Environment, and Health(RPGEN)。

2. 实施概况：自20世纪中期慢性病引起流行病学家的重视以来，前瞻性队列就成为了病因研究的常用设计，如美国Framingham Heart Study、英国British Doctor Study等经典研究。与当今的大型队列相比，这些研究虽然规模稍逊，但其研究设计、实施及项目组织和管理等，时至今日仍颇具借鉴价值，堪称大型队列的先驱。以下仅概述具有代表性的部分现代大型队列。

(1)UK Biobank：被誉为现代大型队列研究的实施范本[1,5-9]。该项目旨在了解英国普通人群中的主要慢性病及其病因，于2007年4月至2010年7月从40～69岁一般人群中招募50万人，完成了多项内容的基线调查(含体格检查)和血、尿样本采集[12-14]。其运行效率和成本控制备受好评，如比预期提前18个月完成了研究对象招募和基线调查，基线费用为1亿美元、维护和随访预计费用约700万美元/年，均在预算范围之内[1]。

UK Biobank的成功得益于多个因素。首先该队列建立于国民卫生服务系统(National Health Service)基础之上。该系统基本覆盖英国全体国民，记录有从出生到死亡的详细健康信息，且有个体唯一编码。UK Biobank从招募研究对象开始就依托此系统，极大提高了可操作性[12,14]。其次，采用集中模式以提高工作效率、节约成本。基线调查时先后设立了36个分中心(assessment center)，最多同时设立6个。平均每个分中心有14名工作人员，每天可完成100名研究对象的基线调查。分中心通过专门的信息传送网络和样本运送通道，每天将采集到的信息和样本发回协调中心，由后者统一保管[14]。此外UK Biobank设计完善，自2000年前后该项目提上议程[15]，至2007年全面启动基线调查，研究团队耗时数年进行了周密的规划和设计，从样本量计算到实验室检测，各环节均有详尽分析，且在正式开始前进行了两次预实验[14]。按原有设计，UK Biobank将通过英国各健康信息系统，以信息关联(linkage)的方式，追踪每名入选研究对象今后数十年的健康结局；并选择有代表性的部分人群，开展暴露水平的重复测量。据UK Biobank网站介绍，目前在曼彻斯特等4个地区已完成了为数2万人的第二轮调查[16]。

(2)NHS：是美国一项迄今已持续近40年的大规模职业人群队列，现已进入第三期研究(NHS Ⅲ)。由于不同时期的研究领导团队不同，其研究内容也有所不同，但总体关注女性生活方式(口服避孕药等)与肿瘤(乳腺癌等)之间的关系，是此类研究中规模最大者之一[17]。

由于研究对象为具有一定医学专业知识的职业人群，因此NHS实现了较高的随访应答率。Ⅰ期队列(12万人)启动于1976年，研究人群为30～55岁已婚女护士，采用邮寄自填式问卷方式，2年进行一次随访；截至2004年，邮寄问卷随访应答率为90%，加上从国家死亡登记系统(US national death index)查找到的死因信息，总体随访率高达98%。Ⅱ期队列(11.6万人)启动于1989年，研究人群为25～42岁女护士，随访方式与Ⅰ期相同，截至目前应答率为90%[18]。Ⅲ期队列启动于2010年，目标研究人群为20～46岁女护士，以网络招募的方式，面向美国和加拿大招募10万志愿者，预计以在线调查的方式、每6个月进行一次随访，目前尚未公布应答率数据[18]。

(3)Millennium Cohort Study(MCS)：美国MCS(英国同名队列研究人群为儿童，可参见http://www.cls.ioe.ac.uk)是有史以来规模最大的军人队列，其目的为研究军旅经历对军人(包括退伍军人)的长期健康影响。该项目预计到2020年前招募20万人，截至目前已达15万人[19]。考虑到研究对象的职业特点(现役军人为主、流动性强等)，MCS在招募研究对象和选择调查模式方面有特殊考虑。

首先，为实现目标样本量，采取了分期招募、逐步追加的方式。截止目前已开展4轮招募，分别为Panel 1(2001—2003年，实际招募77 047人)、Panel 2(2004—2006年，实际招募31 110人)、Panel 3(2007—2008年，实际招募43 440人)和Panel 4(2011年启动，目标招募60 000人)[19]。每轮基线调查后，均开展3年一次的随访。为提高应答率，MCS在招募、基线调查以及随访时，尽量与各种军队纪念日活动相结合，以强化军人的参与荣誉感。截止2009年统计，前三轮的基线调查累计应答率为34%[20](2007年统计为36%[21])。前两轮招募到的10万余人中，超过70%完成了至少一次随访[20]。

其次，在调查模式上，考虑到军队人员流动性较强，但大部分军营配备有计算机及网络等特点，MCS从首轮基线调查开始，就在传统的纸质问卷调查基础上，鼓励研究对象从网络作答，并逐步将网络问卷作为主要的调查、随访模式。结果显示，与纸质问卷调查相比，网络调查能节约50美元/人[22]。

此外，得益于美国军队系统强大的信息体系，MCS与美

国国防部下属的其他诸多信息系统实现了数据链接[23]，由此扩展了本项研究的应用价值。

(4)EPIC：是在欧洲普通人群中研究膳食模式、生活方式、遗传特征与肿瘤等慢性病关系的多中心大型队列研究，由欧洲10国23个研究中心共同参与，总样本量达52万人，覆盖地域广泛、研究人群多样[24,25]。EPIC由国际肿瘤研究协会(International Agency for Research on Cancer, IARC)负责，协调总部设在法国里昂。IARC负责总体的数据保存、生物样本库建立和维护，研究对象招募、基线调查、随访、采样及样本保存等由各参与中心自行开展，由于各参与国的国情和基础不同，包括膳食调查在内的部分研究方法未能完全统一[26]。有资料指出，EPIC对除了膳食以外的其他数据，并未指定统一的数据保存格式，而是由各参与中心自行决定[27]。为保证核心数据(膳食摄入)的质量，研究团队重点评价各国调查工具的有效性和可比性，并开发了一个兼容9种语言的标准化调查软件(EPIC-SOFT)[26]。

(5) Japan Collaborative Cohort Study for Evaluation of Cancer Risk(JACC)：JACC主要关注生活方式与胃癌、肺癌和心血管疾病等的关系，是日本规模最大的人群队列之一[28]。由24个研究机构采取多中心协作模式在45个地区分头开展，基线调查110 792人，定期随访(不同地区频率不同，以查阅医疗档案和入户访视为主)，1988－2009年间的20年总体随访应答率超90%，并完成了抽样重复调查[29]。研究结束后，JACC研究团队总结了项目运行期间的各项失误和教训，特别指出资金来源过于单一、未能建立薪火相传的管理团队，是导致JACC无法持续下去的主要原因，呼吁今后的类似研究应广泛开发和争取稳定的资金来源，并一定要建立起长期、可持续的管理机制[29]。

四、研究特点

与传统队列研究相比，现有大型队列研究具有如下特点：

1. 所研究疾病多为肿瘤、心血管疾病、代谢性疾病等慢性病，尤以发病率低的肿瘤居多。突破了队列研究“不适于发病率很低的疾病病因研究”的传统认识。

2. 样本量巨大(数万甚至数十万)即“超大规模队列(mega cohort)”[30]，是大型队列的最突出特点。庞大的样本量，提高了大型队列在研究低发病率疾病时的研究效能，并使其成为强大的研究平台，诸多子课题可在其中孕育、实施。

3. 随访时间长，如十数年、数十年，甚至终身随访。超长的随访期一方面考虑到研究疾病的潜隐期长，即只有随访足够长时间才能发现病例；另一方面，长期随访也使得研究者有机会直接观察到生命全程中的多种健康结局。但这不可避免的会带来失访增多、组织难度加大、经费难以为继等问题，因此大型队列往往分阶段开展，将目标样本量和相应工作分不同阶段完成。

4. 多建立生物样本库。生物样本库指长期保存研究对象的血、尿、DNA等样本，并将来自同一个个体的样本检测信息、与其暴露和发病等信息进行关联。生物样本库拓展了大型队列研究的应用价值，除支持当前的研究外，还为未来提供了丰富的研究资源，如部分暴露因素可能在研究初期尚未得到充分认识或无法准确检测，但随着时间推移，一旦建立了适宜方法，则可对生物样本库中存储的样本进行检测，从而直接获得研究人群在历史时点的真实暴露水平。

5. 广泛应用现代IT技术。突飞猛进的IT技术在很大程度上改变了流行病学研究的实施方式。在研究对象招募阶段，除传统的寄送邀请函外，越来越多的研究开始尝试采用电子邮件、手机短信、社交媒体以及网络招募等方式。在问卷调查阶段，除纸笔问卷调查和电话调查外，基于便携式设备的计算机辅助调查以及在线调查已逐渐成为主流。并有研究(如UK Biobank)在体格检查中应用计算机程序辅助诊断，可在体检现场快速得出初步结论，有助提高受试者的参与意愿[14]。但IT技术虽然革新了信息获取、传送和存储方式，却对数据安全(尤其是个人隐私信息安全)提出了新的挑战，如何控制新兴调查方式(如在线调查)可能产生的选择偏倚和信息偏倚也有待重视和研究。

6. 注重研究人群的多样性(diversity)。当大型队列的研究目的并非对全人群发病或患病率进行无偏估计，而是揭示疾病与暴露之间的关系时，只要研究涵盖各种特征(年龄、性别、种族、职业、社会经济地位等)人群，即使研究人群对于全人群不具备足够的代表性，在达到必需的样本量后，辅以恰当的统计分析，也足以揭示不同人群中疾病与暴露之间的关系。基于此，出于节省成本的考虑，在一般人群中建立大型队列通常会放弃对于高应答率(high response rate)的追求(即致力于让每一个目标研究对象都参与调查，以提高研究人群对于全人群的代表性)，而改为着力保证入选研究对象的多样性，即“高招募率(high recruitment rate)”[1]。

7. 越来越多的以信息关联的方式获取结局信息。传统的前瞻性研究一般通过定期主动随访，从研究对象直接收集发病、死亡信息。此种方式虽然准确度较高，但对于超大规模队列而言，组织、实施难度极大。因此在有条件的地区，信息关联逐步成为了获取结局信息的主要方式。所谓信息关联，指使用个体唯一性标识(身份证号、医疗或社会保险号、驾照证号等)，从疾病登记、监测等卫生信息系统，门(急)诊、出入院等医院信息系统，以及医疗保险赔付等健康相关信息系统中提取研究对象的发病、就诊、死亡信息。UK Biobank以及我国CKB均已采取此种方式进行结局追踪[12,31]。

8. 高度重视成本控制。成本控制是大型队列研究最重要的考虑因素，堪称决定成败的关键[1,5-8]。上述集中模式、IT技术、信息关联等均可视为成本控制措施。在建立生物样本库的过程中应用工业化流程，其目的之一也是控制成本。

五、争论及展望

大型队列是人类社会科技发展和物质资源积累到一定程度的产物，它的出现无疑开启了现代病因研究的新篇章。在各基金机构的支持下，顺应大数据时代的趋势，大型队列日渐获得关注，争论也随之产生，例如对于罕见疾病，超大规模队列是否是解决样本量不足的唯一或最佳选择？大型队列是否真的能做到兼顾数据的数量和质量？在资源总量有

限的前提下，大型队列对其他类型的流行病学研究将带来怎样的影响？以及流行病学研究者应该如何平衡“大而强”(bigger is better)和“小而精”(small is beautiful)[8]？虽然大型队列在病因学研究中的价值目前面临以上质疑，但其作为强大研究平台的实用意义已得到了一致肯定。可以预期，在今后一段时期，大型队列仍将持续发展，随着时间的推移，其对流行病学，乃至医学研究的真正意义必将逐步展现。

参 考 文 献

[1] Manolio TA, Collins R. Enhancing the feasibility of large cohort studies[J]. JAMA, 2010, 304(20): 2290-2291.

[2] Chakravarti A, Little P. Nature, nurture and human disease[J]. Nature, 2003, 421: 412-414.

[3] World Health Organization. Global status report on noncommunicable disease 2010[R]. Geneva: WHO, 2011.

[4] Abegunde DO, Mathers CD, Adam T, et al. The burden and costs of chronic diseases in low-income and middle-income countries [J]. Lancet, 2007, 370: 1929-1938.

[5] Gaziano JM. The evolution of population science: advent of the mega cohort[J]. JAMA, 2010, 304(20): 2288-2289.

[6] Collins FS. The case for a US prospective cohort study of genes and environment[J]. Nature, 2004, 429: 475-477.

[7] Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies[J]. Nat Rev Genet, 2006, 7(10): 812-820.

[8] Hense HW. When size matters[J]. Int J Epidemiol, 2011, 40(1): 5-7.

[9] Manolio TA, Weis BK, Cowie CC, et al. New models for large prospective studies: is there a better way?[J]. Am J Epidemiol, 2012, 175(9): 859-866.

[10] Public Population Project in Genomics and Society. Study catalogue [EB/OL]. [2013-11-04]. http://www.p3gobservatory.org/studylist.htm; jsessionid=9F34E079CB3C6A8FA7383350023B5730.

[11] Li L, Guo Y, Chen Z, et al. Epidemiology and the control of chronic disease in China, with emphasis on the Chinese Biobank Study[J]. Public Health, 2012, 126(3): 210-213.

[12] Collins R. What makes UK Biobank special?[J]. Lancet, 2012, 379: 1173-1174.

[13] Allen N, Sudlow C, Downey P, et al. UK Biobank: current status and what it means for epidemiology[J]. Health Policy Technol, 2012, 1(3): 123-126.

[14] UK Biobank Coordinating Center. UK Biobank: protocol for a large-scale prospective epidemiological resource [R/OL]. [2013-11-04]. http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf? phpMyAdmin=trmKQlYdjnQIgJ%2CfAzikMhEnx6.

[15] Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality[J]. Pharmacogenomics, 2005, 6(6): 639-646.

[16] UK Biobank. 20000 participants return for a repeat assessment [EB/OL]. [2013-11-04]. http://www.ukbiobank.ac.uk/2013/09/20000-participants-return-for-a-repeat-assessment/.

[17] Colditz GA, Hankinson SE. The nurses' health study: lifestyle and health among women[J]. Nat Rev Cancer, 2005, 5(5): 388-396.

[18] Nurse's Health Study. History [EB/OL]. [2013-11-04]. http://www.channing.harvard.edu/nhs/? page_id=70.

[19] The Millennium Cohort Study. About the study[EB/OL]. [2013-11-04]. http://www.millenniumcohort.org/aboutstudy.php.

[20] Smith TC. The US Department of Defense Millennium Cohort Study: career span and beyond longitudinal follow-up [J]. J Occup Environ Med, 2009, 51(10): 1193-1200.

[21] Ryan MAK, Smith TC, Smith B, et al. Millennium cohort: enrollment begins a 21-year contribution to understanding the impact of military service[J]. J Clin Epidemiol, 2007, 60(2): 181-191.

[22] Smith B, Smith TC, Gray GC, et al. When epidemiology meets the internet: web-based surveys in the Millennium Cohort Study [J]. Am J Epidemiol, 2007, 166(11): 1345-1354.

[23] Smith TC. Linking exposure and health outcomes to a large population-based longitudinal study: the Millennium Cohort Study[J]. Mil Med, 2011, 176: 56-61.

[24] Bingham S, Riboli E. Diet and cancer-the European Prospective Investigation into Cancer and Nutrition [J]. Nat Rev Cancer, 2004, 4(3): 206-215.

[25] International Agency for Research on Cancer. About EPIC [EB/OL]. [2013-11-04]. http://epic.iarc.fr/research/meth.php.

[26] International Agency for Research on Cancer. Research activities: methodological issues[EB/OL]. [2013-11-04]. http://epic.iarc.fr/research/meth.php.

[27] Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition(EPIC): study populations and data collection [J]. Public Health Nutr, 2002, 5(6B): 1113-1124.

[28] Tamakoshi A. Overview of the Japan Collaborative Cohort Study for Evaluation of Cancer(JACC)[J]. Asian Pac J Cancer Prev, 2007, 8 Suppl: 1-8.

[29] Tamakoshi A, Ozasa K, Fujino Y, et al. Cohort profile of the Japan Collaborative Cohort Study at final follow-up [J]. J Epidemiol, 2013, 23(3): 227-232.

[30] Sun DJY, Lv J, Li LM. Mega cohort: a powerful tool for etiologic research on complex human diseases in 21st century [J]. Chin J Dis Control Prev, 2013, 17(1): 66-70. (in Chinese)
孙点剑一，吕筠，李立明. 流行病学超大规模队列研究——开启21世纪人类复杂性疾病病因研究的钥匙[J]. 中华疾病控制杂志，2013，17(1)：66-70.

[31] 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. (in Chinese)
李立明，吕筠，郭彧，等. 中国慢性病前瞻性研究：研究方法及调查对象基线特征[J]. 中华流行病学杂志，2012，33(3)：249-255.

（收稿日期：2013-11-07）
（本文编辑：张林东）