大型人群队列研究

王笑峰, 金力*

复旦大学生命科学学院, 现代人类学教育部重点实验室流行病学室, 上海 200438

* 联系人, E-mail: lijn@fudan.edu.cn

收稿日期: 2016-01-10; 接受日期: 2016-02-11

科技部国际科技合作专项(批准号: 2014DFA32830)和国家科技支撑计划重点项目(批准号: 2011BAI09B02)资助

摘要 在精准医学和大数据时代的今天, 大型人群队列研究成为流行病学研究的主旋律之一. 人群队列研究由传统小样本队列发展到了当前的大型队列, 而构建大型人群队列的模式也在不断地演化. 各组学、大数据科学、分子影像等新技术不断被纳入到人群队列研究中以回答不断产生的广泛科学问题, 服务于疾病精准的预防、诊断和治疗. 本文将重点讨论大型人群队列建设的时代背景、必要性、建设现状和新模式、数据分析等, 并提出我国建设大型队列所面临的挑战和机遇.

关键词 大型队列, 人群, 建设模式, 精准医学

人群队列研究通过对一定人群进行随访和纵向观察, 评估暴露和终点的关系^[1]. 队列研究可揭示疾病的病因、评价预防效果、揭示疾病的自然史、掌握人口健康状况、引导实验设计, 将知识转化为临床和人群早期诊断和干预策略^[1], 从而提高疾病的防治水平, 降低社会卫生负担. 为此, 历史上建立了很多传统队列, 例如, 美国的弗莱明翰队列 Framingham Study^[2](每代约五千余人, 几十年间做了 3 代人)及其遗产队列 Atherosclerosis risk in communities study^[3](ARIC, 1.6 万例), Cardiovascular health study^[4](CHS, 约 6 千例), 荷兰的 Rotterdam study^[5](约 8 千例), 丹麦的 Copenhagen city heart study^[6](约 1 万例).

这些样本量较小的传统队列人群常见重大慢性病的防控作出了巨大的贡献. 其中, 从 1947 年起在美国一个小镇上开展的弗莱明翰队列^[2]产生了一

系列划时代的发现, 例如, 发现高血压、高胆固醇、吸烟增加心脑血管病的风险, 构建了冠心病的发病风险预测模型, 提出心力衰竭诊断标准等. 这些发现普适于不同地域、不同时代的人群. 对这些危险因素的干预最终导致了全球范围内的心脑血管病发病和死亡的剧烈下降.

1 小样本队列的不足

然而, 很多传统小样本人群队列存在以下缺点:

(i) 样本量不足导致效应估计不稳定, 表型测量也不见得很多, 难以开展表型组研究. 潜在风险测量不完整而导致疾病关联的系统性低估, 混杂因素测量不完全而导致效应的高估或者低估.

(ii) 很多已建的小样本传统队列往往不能代表

引用格式: 王笑峰, 金力. 大型人群队列研究. 中国科学: 生命科学, 2016, 46: 406–412

Wang X F, Jin L. Large population-based cohort studies. Sci Sin Vitae, 2016, 46: 406–412, doi: 10.1360/N052016-00104

一般人群, 在人群多样性、暴露和疾病的种类、方法的标准化等方面不足. 尤其是对少数民族、弱势人群, 低年龄组等人群的代表性不足^[7,8]. 这些队列常常为了增加数据质量或强调环境暴露间的比较而未考虑样本对源人群的代表性.

(iii) 一些队列设计和操作不理想, 难以防止偏倚, 而利用传统队列做研究时有时对照可能被病例污染, 临床相关的混杂因素有时也不能被系统鉴定, 导致弱队列(weak cohort)的产生.

(iv) 队列的相关研究设施花费很大, 包括表型注释和样品采集. 尤其是将高新技术用在这些传统已建队列去测量新兴风险因素时的经费开支异常大, 操作难度也较大.

(v) 在当前的研究者和知情同意的前提下, 传统小队列数据开放共享的流畅性有时也显得不足.

2 进行大型队列研究的必要性

目前, 大型队列尚无严格定义. 其样本量可为几万以上, 而目前的规模常为十几万以上或几十万例个体. 将疾病特异性的小队列简单合并做成大型队列, 无论从效率和资金上均是不可取的, 这就需要创新的队列建设模式, 其中资金是首先需要考虑的因素, 而专业化高效操作同科学性一样重要^[9]. 近年来, 由于传统队列的不足, 技术的进步、科学和医疗大设施的建立等原因, 传统队列受到数十万规模的大型队列的挑战.

为什么要建设大型队列? 复杂疾病的发生是基于生活方式、环境、遗传的复杂相互作用所致. 然而, 纵然经过了多年研究, 这些因素导致不同疾病的清晰图案依然不清, 需要建设大型人群队列来精细解析疾病的病因结构. 这主要基于检出阳性的统计效率考虑. 例如, 表 1 以 2000 年美国调查的结直肠癌 5 年随访的数据为例, 20 万人的队列在随访结束时理论

上仅能收集 456 例病例, 而 50 万人的队列能收集 1141 例病例, 而 100 万人的队列就可累积 2282 例病例, 此时就有较大的新发样本通过队列内的巢式病例对照设计进行结直肠癌的病因学研究. 此时如果还想进行遗传变异同环境因素的交互作用分析, 就需要随访更长时间才能获得有足够统计效率的样本量.

既往在做干预实验时就有项目将观察性队列结合干预性队列一起进行操作. 一个成功的例子是美国国会资助美国国立卫生研究院 (National Institutes of Health, NIH) 开展的 WHI(women's health initiative)^[11]. 它包含 3 个重叠的临床实验(针对心血管病、癌症和骨质疏松, 干预实验样本量 6.45 万)及一个观察性队列(10 万例), 研究结果颠覆了一些既往的观念, 例如, WHI 不支持很多观察性研究发现的绝经女性使用雌激素替代疗法来预防心血管病发病.

在当前的精准医学时代, 因为募集足够量的实验对象越来越困难, 新药开发的成本越来越高. 埋藏干预性研究在观察性大型队列中将会是一种比较快速且经济有效的选择. 例如, 在 100 万规模的精准医学计划队列中, 在队列建成的 3~5 年后即可高成本效益地开展药物临床实验(表 2). 而在如此大样本量的队列中, 其他类型的研究, 如探索常染色体隐性遗传病的杂合失功能突变的效应研究在队列建成后的 3 年内即可以开展.

3 大型人群队列的建设现状

面对传统队列的缺点, 为了定量评估不同慢性病的效应, 需要发展新的策略去构建一些大型的人群为基础的前瞻性队列研究. 一般情况下, 全基因组关联研究(genome-wide association study, GWAS)鉴定的复杂性状及常见病的效应多在风险比 1.3~1.5 以内.

表 1 前瞻性队列研究的发病数估计^[10]

发病数(率%)/10 万人 年 ⁻¹	疾病	不同样本量队列 5 年后的发病数		
		200000	500000	1000000
10(0.01)	帕金森氏病、精神分裂症	91	228	457
50(0.05)	结肠直肠癌、肾衰竭	456	1141	2282
100(0.10)	乳腺癌、髌部骨折	912	2279	4559
200(0.20)	糖尿病、卒中、心力衰竭	1820	4550	9100
500(0.50)	心肌梗死、癌症	4524	11309	22618
3000(3.00)	白内障、高血压	25858	64644	129289

表 2 精准医学队列预计实现不同种类研究的时间表(每格的+代表估计进行不同类型研究的时间表)^[12]

队列能力		时间(年)			
		0~2	3~5	5~10	>10
队 列 效 能	(1) 发现疾病风险因素	+	+++	+++	++++
	(2) 药物基因组学	+	+++	+++	+++
	(3) 生物标志物	+	++	+++	+++
	(4) 移动健康同疾病结局的关联		+	++	++++
	(5) 失功能突变的效应		+	+++	+++
	(6) 疾病新分类		+	+++	++++
	(7) 授权数据的个体	+++	+++	+++	+++
	(8) 靶向治疗的临床实验		+	+++	+++

由于传统小样本队列样本量的不足,其建设者们往往组成队列联盟(cohort consortium)来鉴定各种表型的遗传易感性位点,例如,CARDIoGRAM plusC4D 联盟整合了一系列队列的 6.3 万例冠心病病例及 13 万对照,解析了冠心病的遗传结构^[13]。然而队列联盟有以下缺点:组成队列联盟的原始队列间的操作流程往往不统一,从而影响队列间数据的有效合并。环境暴露在改变,技术也在更新,这些新信息在已建的队列联盟中无法获得。另一种尝试是电子病历聚合式大型队列,例如,电子医疗记录和基因组网络(electronic medical records and genomics network, eMERGE)。eMERGE 通过网络获取 5 个美国成员医疗集团日常临床诊治获取的数据,即能有足够的准确度进行 GWAS 研究。例如,eMERGE 用日常临床诊治的数据进行表型组研究,不但验证了既往 GWAS 已鉴定的 66%(51/77)的阳性结果,并且发现了 63 个新的多效性基因-表型关联^[14]。

除了以上队列建设的尝试外,分散式大型队列及集中式大型队列为最常见的队列建设模式。

3.1 分散式大型队列

分散式大型队列常由大型医疗中心们组建多地委员会,每个中心募集和随访本医疗集团覆盖的调查对象。分散式模式可测量较细致的表型,且可以由课题组长们通过不同途径改善技术和方法并评价发现的重现性。然而分散性队列需要很昂贵的半永久性研究中心去维持,且标准化也往往受到挑战。

欧洲营养与肿瘤前瞻性调查(The European Prospective Investigation into Cancer and Nutrition, EPIC)是一个经典的分散式大型队列。EPIC 成员为欧洲 10 国 23 个中心,其数据为中心化存储。EPIC 运行的另一个层面是成立各研究工作组,如 EPIC-heart 工

作组。EPIC-heart 是多中心前瞻性使用信息化数据库和生物样本调查营养等生活方式同心血管病发病的关系。随访通过死因登记、主动随访和死亡记录收集^[15]。EPIC 的随访通过各成员国每 3~4 年重新接触调查对象获得生活方式和疾病的变化。在肿瘤发病和总死亡随访方面,每个肿瘤监测点都建立一个工作组(终点委员会)来收集和标准化临床和病理数据。

3.2 集中式大型队列

集中式大型队列将大多队列建设活动集中于一个或几个协调中心,进行中心化操作。协调中心的活动有选择评估中心、仪器装备、人员、供给、邀请调查对象、应答、募集、重新接触调查对象、存储库建设等。而各个评估中心聚焦于基线调查、传输标本和数据。为了节省成本,评估中心在样本收集完毕后就关闭。

英国的国家队列 Uk Biobank(UKB)是一个经典的集中式大型队列。UKB 目的是调查复杂病的环境和遗传因素,2007~2010 年募集了 50 万余例 40~69 岁个体,进行了 2~3 h 的基线评估。中心化的组织模式提高了募集效率,在预算开支内(仅约 1 亿美元)就提前完成了 50.3 万人的募集。目前他们还在进行 10 万人的体力活动检测,80 万个 SNPs 检测,10 万人的脑部、胸部核磁共振、颈部超声和全身骨密度测量^[16]。

4 大型人群队列构建的新模式——精准医学计划队列

经过十多年来组学技术、数据收集和存储技术的进步和移动健康(mHealth)技术的长足发展,创建一个大规模精准医学队列遇到了一个前所未有的机遇。2015 年初,美国总统适时启动“精准医学计

划”(precision medicine initiative, PMI)^[17]. 精准医学通过更精确的测量影响健康和疾病的分子、环境和行为因素来重新定义疾病发生、预后、治疗反应、健康结局, 从而产生更精准的诊断、更合理的疾病预防策略和更好的治疗选择, 并开发新型治疗方案.

为了完成此计划, 美国启动了精准医学队列项目 (precision medicine initiative cohort program, PMI-CP)^[12]来提供一个平台服务于精准医学, 推动对健康和疾病的理解并为用新方式做研究打基础, 从而在未来的多年里受益于美国人民, 并保障美国在领域内的竞争力.

2016 年政府为精准医学队列投入 1.3 亿美元. 此队列预计 4 年募集 100 万以上覆盖美国全人群(不同年龄段)的研究对象, 至少随访 10 年来获得足够新发病例以及环境和治疗暴露.

精准医学计划队列咨询委员会工作组计划 PMI 队列成为一个志愿者参与的新建、广泛共享的国家研究资源. 工作组识别了两种不同的募集方式, 直接志愿者和医疗供给组织(Healthcare Provider Organization, HPO)合作收集(HPO 收集时需考虑队列能反映人群多样性、电子健康记录的效率、随访的时限). 数据收集包含人口统计学、自报健康、行为和生活方式测量、传感为基础的数据(通过手机、穿戴、居家为基础的设备)、电子健康记录获得的结构性临床资料、基线健康检查数据、医疗相关数据(如门诊药)、疾病特异性的监测(在亚组队列中)、社会网络、空间和环境数据、标本相关实验室数据(基因组、蛋白组、代谢组、单细胞研究、感染暴露、临床生化、组织病理等).

为了促进资料接触、标准化和参与者加入, PMI 队列采用“轮辐模型”进行操作, 包含一个协调中心保障协调数据和样本的协调、研究对象的交流和参与, 以及研究的开展. PMI 工作组鼓励 NIH 同非盈利和商业组织进行新型合作来获得最先进的分析方法、弹性存储和分析, 以及技术能力.

为了保障资料的存储和接触, 工作组推荐采用“杂交模型”的数据结构和分析结构, 利用中央存储的核心数据集而同时保障联邦(各 HPO)在网络节点上接触附加数据. PMI 队列的生物银行科保存血液、微生物标本、指甲头发等标本. 队列的管理为队列项目主任负责制, 工作组推荐的管理结构包含一个各利益相关者组成的管理委员会, 它们被一个小一些

的执委会领导. 5 个亚委员(结果和信息返回亚委员会、数据亚委员会、资源接触亚委员会、生物银行亚委员会、安全亚委员会)将向执委会和 PMI 队列主任汇报工作.

5 大型队列的数据分析

人群队列结局事件发生率的计算是队列研究数据分析的关键. 频率的描述性测量指标常用发病率和死亡率. 标化发病比(standardized incidence ratios, SIR)和标化死亡比(standardized mortality ratios, SMR)用来同一般人群比较发病率和死亡率^[18]. 效应测量指标常用相对危险度(relative risk, RR)和归因危险度(特异危险度、率差, rate difference, RD). 回归分析调查潜在风险因素对靶指标的影响, 其中 Cox 回归使用风险比例模型来计算风险比(hazard ratio, HR). 虽然 Logistic 回归对于两分类变量病例对照设计是计算风险比的最佳模型, 它也可以被用于队列研究(如当结局在设定的时间点获得)^[18].

如果发生终点事件的时间被记录, 常使用生存分析模型, 如非参数的 Kaplan-meier 法和半参数的 Cox 比例风险模型来分析终点. 生存分析方法采用删失技术容纳个体的损耗, 可聚焦具有临床显著意义的终点并使用所有可获得的数据^[19]. 生存模型临床有效性的一个关键测量是区分那些将发生感兴趣疾病与不发生疾病的个体. 各种测量方式被提出用以捕捉辨识度, 而受试者工作曲线(receiver operating characteristic curve, ROC)下面积(area under curve, AUC)是最流行的度量. 除了用 AUC 外, 常用净重新分类指数(net reclassification improvement, NRI)来评价新型标志物对终点的附加预测能力^[20].

人群队列随访产生的纵向数据在理论和实践上往往有以下问题: 个体终点重复测量间具有相关性, 测量的时间间隔不等距, 随机和非随机性数据缺失, 协变量常随着随访时间而变化(如年龄、体重). 广义线性模型中的混合效应模型和广义估计方程(generalized estimating equation, GEE)在处理这类复杂纵向数据中发挥着重要的作用^[21]. 混合效应模型和 GEE 可选择不同的模型假设来灵活扩展数据的处理方式, 各种可及的数据都可被使用, 可用于连续、等级、计数或分类数据. 可容许缺失数据, 进行协变

量调整, 包括调整基线的严重水平^[19].

队列研究能在疾病发生前测量环境暴露, 因此比病例对照设计更有效研究遗传环境相互作用. Logistic 或 Cox 回归允许单位点基因型和复杂的暴露数据灵活拟合并调整多种混杂因素, 而高维基因-基因或基因-环境互作最常用多因子降维法(multifactor dimension reduction, MDR)进行分析^[22]. 使用生存分析技术可研究遗传及基因-环境互作对疾病进展及有删失数据事件的效应, 使用混合模型及 GEE 可分析其对纵向表型的效应^[23].

6 我国当下进行人群大型队列研究面临的挑战与机遇

经过多年的建设, 我国已建成了若干大型队列, 例如, 规模为 50 万样本量的中国慢性病前瞻性研究项目^[24]、规模为 20 万例的泰州人群健康跟踪调查^[25], 规模为 18 万例的中国高血压随访调查队列^[26]、规模为 7.5 万例的上海女子队列^[27]等队列. 这些队列将为我国生物医学研究及疾病的防控做出重要贡献. 然而, 在此精准医学时代里, 中国进行人群大型队列研究面临着一系列挑战与机遇.

6.1 挑战

虽然严格评估的新型预防和治疗策略已经导致人群整体层面的发病和死亡率的下降. 然而, 未识别的疾病病理、环境暴露、社会和行为及遗传因素导致的个体间变异导致很多人药物无效甚至差的预后. 目前, 因为个体病因和治疗反应变异的知识不完整导致对每个个体进行最佳疗效鉴定的进展较弱. 另外, 理想途径首先应该能防得住疾病的进展. 这需要识别特定疾病的高危个体从而开发新型干预来预防后续疾病的进展. 如何在大型人群队列研究中鉴定相关因素, 开发相关干预策略是我国大型队列研究面临的一个挑战.

医疗记录的综合收集, 不同的医疗系统资源高效共享、交换和合并, 以及病人能否接触到其自身的数据一直是大型队列研究的一个重要挑战. 然而, 链接电子健康档案的大型队列网络不仅有潜力加速生物医学研究, 也可在纵向研究中被随访个体的健康.

各组学、分子和影像研究的巨大开支、缺乏强有力的移动数据测量能力影响了队列的规模和研究的

精细化. 然而, 目前这些技术已经能够经济有效地应用到人群队列研究中来.

队列需要反应人群的多样性, 尤其是对弱势群体(如儿童、认知能力受损的个体)的代表性不足一直是既往队列研究受诟病的一点. 建设宽泛的反应国人多样性的队列(尤其是在某些亚层中进行过抽样)将应对这个挑战.

队列建设既往一个重要的挑战是如何使公众得知其收益而保障其参与性. 然而, 同科学技术进步同时发生的是医疗实践和医学研究文化的改变. 如今, 医生、研究者、病人、社区越来越多参与讨论如何提升个体、家人和社区的健康. 参与者目前已经不仅是病人或研究对象, 而逐渐变成了伙伴, 这降低了研究对象的募集、长期随访监测以及邀请他们参加随后临床实验的难度.

6.2 机遇

虽然有诸多挑战, 我国在建设大型队列方面也面临着前所未有的机遇.

既往在中国建设大型队列难度最大的是队列的长期随访, 而目前医院电子病历的普及, 社区电子健康档案的普遍建档, 重点人群的公共卫生服务措施(新生儿访视及儿童保健管理、孕产妇访视、65 岁以上老年人定期健康体检), 高血压及精神病的登记管理和定期随访为队列人群的疾病发病的长期随访监测提供了新的渠道.

移动健康技术的飞速进展, 手机、可穿戴式和居家传感设备的普及使人们能够支付得起进行日常活动、行为、心血管等功能等方面的监测. 具有临床意义的革新设备如心理健康测量设备也被快速设计出来为人群普遍使用. 传感设备已经越来越多地被用于检测身体指标的变化同疾病的关系. 全球定位系统(global positioning system, GPS)可以整合到环境污染和日常活动中来进行环境暴露的测量. 这些价廉的主动持续监测设备为队列评估人群健康提供了新的测量手段. 营养和体力活动的精准测量也得到了长足的改善.

近期生物医学技术的飞速发展, 为低价基因组测序、代谢组、信号分子免疫系统的活性测定提供了新的机遇. 大数据分析技术应用到医学影像学可更加精准地对疾病谱和疾病的特殊发展阶段进行分类.

7 结语

在这个精准医学和大数据的时代里, 队列研究已经成为了流行病学的主旋律之一^[28]. 在国家持续的投入和长效科研机制的保障下, 发挥我国行政约束力强, 整合医疗卫生资源和构建网络的制度优势. 抓住各组学技术、各种分子表型测量技术日益价廉, 移动健康和医疗等新技术进入家庭的契机, 构建反

应国人多样性的国家级人群大型健康队列. 并且通过链接电子医疗记录进行长期随访监测, 以及加强同医疗系统、参与者和公众互动以促进理解、共享和转化. 以上措施将有助于推动国人大型队列研究, 进而建立多层次精准医疗知识体系和生物医学大数据平台, 形成重大疾病风险评估和预测、早筛分类, 个体化治疗及疗效监测的防治方案.

参考文献

- 1 Sorlie P, Wei G S. Population-based cohort studies: still relevant? *J Am Coll Cardiol*, 2011, 58: 2010–2013
- 2 Wong N D, Levy D. Legacy of the framingham heart study: rationale, design, initial findings and implications. 2013, *Global Heart*, 8: 3–9
- 3 The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am J Epidemiol*, 1989, 129: 687–702
- 4 Fried L P, Borhani N O, Enright P, et al. The Cardiovascular Health Study: design and rationale. *Ann Epidemiol*, 1991, 1: 263–276
- 5 Hofman A, Brusselle G G, Darwish Murad S, et al. The Rotterdam Study: 2016 objectives and design update. *Eur J Epidemiol*, 2015, 30: 661–708
- 6 Appleyard M, Hansen A T, Schnohr P, et al. The Copenhagen City Heart Study. A book of tables with data from the first examination (1976–1978) and a five-year follow-up (1981–1983). *Scand J Soc Med*, 1989, 170: 1–160
- 7 Willett W C, Blot W J, Colditz G A, et al. Merging and emerging cohorts: not worth the wait. *Nature*, 2007, 445: 257–258
- 8 Collins F S, Manolio T A. Merging and emerging cohorts: necessary but not sufficient. *Nature*, 2007, 445: 259
- 9 Manolio T A, Weis B K, Cowie C C, et al. New models for large prospective studies: is there a better way? *Am J Epidemiol*, 2012, 175: 859–866
- 10 Manolio T A, Bailey-Wilson J E, Collins F S. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet*, 2006, 7: 812–820
- 11 Nabel E G. The Women's Health Initiative—a victory for women and their health. *JAMA*, 2013, 310: 1349–1350
- 12 Precision Medicine Initiative (PMI) working group report to the advisory committee to the director, NIH. The Precision Medicine Initiative Cohort Program—Building a Research Foundation for 21st Century Medicine. 2015, September 17.
- 13 CARDIoGRAMplusC4D consortium. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet*, 2013, 45: 25–33
- 14 Denny J C, Bastarache L, Ritchie M D, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*, 2013, 31: 1102–1110
- 15 Riboli E, Hunt K J, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr*, 2002, 5: 1113–1124
- 16 Ganna A, Ingelsson E. 5 year mortality predictors in 498,103 UK Biobank participants: a prospective population-based study. *Lancet*, 2015, 386: 533–540
- 17 Remarks by the President in State of the Union Address. 2015-01-20, <https://www.whitehouse.gov/the-press-office/2015/01/20/remarks-president-state-union-address-january-20-2015>
- 18 Ressing M, Blettner M, Klug S J. Data analysis of epidemiological studies: part 11 of a series on evaluation of scientific publications. *Dtsch Arztebl Int*, 2010, 107: 187–192
- 19 Hennen J. Statistical methods for longitudinal research on bipolar disorders. *Bipolar Disord*, 2003, 5: 156–168
- 20 Pencina M J, D'Agostino R B Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*, 2008, 27: 157–172
- 21 Edwards L J. Modern statistical techniques for the analysis of longitudinal data in biomedical research. *Pediatr Pulmonol*, 2000, 30: 330–344
- 22 Thomas D. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annu Rev Public Health*, 2010, 31: 21–36

- 23 Liu C Y, Maity A, Lin X, et al. Design and analysis issues in gene and environment studies. *Environ Health*, 2012, 11: 93
- 24 Chen Z, Lee L, Chen J, et al. Cohort profile: the Kadoorie Study of Chronic Disease in China (KSCDC). *Int J Epidemiol*, 2005, 34: 1243–1249
- 25 Wang X, Lu M, Qian J, et al. Rationales, design and recruitment of the Taizhou Longitudinal Study. *BMC Public Health*, 2009, 9: 223
- 26 Gu D, Kelly T N, Wu X, et al. Mortality attributable to smoking in China. *N Engl J Med*, 2009, 360: 150–159
- 27 The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am J Epidemiol*, 2005, 162: 1123–1131
- 28 Khoury M J. Planning for the future of epidemiology in the era of big data and precision medicine. *Am J Epidemiol*, 2015, 182: 977–979

Large Population-Based Cohort Studies

WANG XiaoFeng & JIN Li

Epidemiology Unit of Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200438, China

In the era of big data and precise medicine, the construction of large cohorts is a main theme in epidemiological studies. From small traditional cohorts to current large cohorts, the construction model of population-based cohorts is constantly evolving. New technologies such as omics, big data science, and molecular imaging has contributed to cohort studies to resolve a wide range of scientific problems and to enable accurate prevention, diagnosis, and treatment of diseases. This review will focus on the background, necessity, status and new models, data analysis for constructing large cohorts, and challenges and opportunities for China to construct large cohorts.

large cohorts, population, construction models, precision medicine

doi: 10.1360/N052016-00104