◇综　　述◇

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

孙点剑一，吕筠，李立明

【摘要】 随着现代病因理论对于人类常见疾病“遗传-环境交互作用”的共识，流行病学超大规模队列研究在近20年来得到了快速发展，并逐渐成为21世纪开展人类复杂性疾病病因研究最有利的工具和平台之一，但与此同时，超大规模所带来的执行管理、成本控制和资源配置等多方面的挑战也不容忽视。

【关键词】 队列研究；患病率；危险因素

【中图分类号】R181.23；R821.35　　【文献标识码】A　　【文章编号】1674-3679(2013)01-0066-06

Mega cohort: a powerful tool for etiologic research on complex human diseases in 21st century

SUN Dian-jian-yi, LV Jun, LI Li-ming. *Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing* 100191, *China*

【**Abstract**】 In light of the modern etiological consensus on gene-environment interactions that contribute to human common diseases, great progress has been made during the last two decades in mega cohort studies. Thus, it has gradually become one of the most powerful tools and useful platforms for etiologic research on complex human diseases in 21st century. However, it is much more attention that we should pay to the big challenges brought by the mega sample size, such as execution, management, cost control, resources optimization, etc.

【**Key words**】 Cohort study; Prevalence rate; Hazard

(*Chin J Dis Control Prev* 2013, **17**(1): 66-71)

复杂性疾病(complex disease)，于1784年完整出现在医生Ware[1]的病例记录中，但特指患者症状和体征的复杂性。直到20世纪60～80年代，随着一系列疾病概念的提出[2-7]，复杂性疾病的意义才逐渐由临床特征转变到病因机制上来。20世纪80年代末期，病因研究迈入微观层面，人类复杂性疾病(complex human disease)一词开始在遗传学领域被提出。20世纪90年代初至今，随着人类基因组计划(human genome project，HGP)的完成和对于基因功能的不断认识，以及基因-环境交互作用(gene-environment interactions)和表观遗传学(epigenetics)的提出和发展，研究者发现大部分人类常见疾病并不是单纯基于环境暴露或遗传变异而发生的，而更可能是基因与环境的交互作用所致[8]，所以最终形成了“基因、环境、基因-基因、环境-环境、基因-环境交互作用”为现代病因理论基础的人类复杂性疾病这一概念。

【基金项目】“十二五”国家科技支撑计划(2011BAI09B01)
【作者单位】北京大学公共卫生学院流行病与卫生统计学系，北京　100191
【作者简介】孙点剑一(1987－)，男，湖北荆州人，在读博士研究生。主要研究方向：慢性流行病学。
【通讯作者】李立明，E-mail: lmlee@vip.163.com

人类复杂性疾病主要以慢性非传染性疾病和心理、精神类疾病为主，具有病程长、预后差和致残高的生物医学特点，以及负担重、影响广和耗资大的社会经济学特点[9-13]，已成为并将长期成为21世纪威胁人类健康的首要原因[14]。此外，基于全民预防和个体化治疗的需要，人类复杂性疾病的病因研究也将注定是21世纪生物医学领域的核心。

1　流行病学超大规模队列研究的产生

20世纪后半期，基于人类复杂性疾病(特别是恶性肿瘤)病因研究的客观需求，最初的一批流行病学超大规模队列研究(以下简称为Mega Cohort)诞生，如美国的Nurses' Health Study[15]、日本的The JACC Study[16]以及欧洲的EPIC研究[17]，样本规模均超过10万。Mega Cohort最重要的特点即是超大的队列样本规模，从传统流行病学的理论出发，大规模队列研究设计是针对多结局最好的观察性研究方法[18]，统计分析把握度更高，可重复性更强，并可根据不同的结局和暴露进行选择性抽样以开展嵌入式研究[19]。然而，从当今研究的实际出发，Mega Cohort的产生却是基于21世纪人类复杂性疾病病因研究的客观需求，主要体现在以下两点：

第一，人类复杂性疾病，潜隐期长，共享一定的危险因素（如吸烟、肥胖和少体力活动等），不仅包括高血压、白内障等发病率或患病率较高的一类常见疾病，也包括帕金森氏病、精神分裂症等发病率或患病率较低的一类罕见疾病。对此，Mega Cohort 研究以其超大的样本规模在提高研究效能的同时，也能弥补传统队列研究在罕见疾病研究中难以获取病例或获取足够病例的劣势。如表 1 所示，以帕金森氏病和/或精神分裂症为例，一个随访 10 年样本量为 5 000的队列在结束时理论上仅能发现 4 例，但一个随访 10 年样本量为 50 万的队列研究却能累积 424 例。此时不仅可以通过传统队列研究设计进行效应估计，还能通过衍生病例对照研究设计（如巢式病例对照研究、病例队列研究和两阶段病例对照研究[20]等）的方式计算效应值。

第二，人类复杂性疾病病因机制复杂，其发生和流行往往是多个危险因素综合作用的结果，如 2002 年美国 Framingham Heart Study 证实了多种危险因素对高血压病发生的协同作用[21]以及脂肪摄入与肝脂肪酶缺陷（hepatic lipase deficiency，LIPC）基因型对于高密度脂蛋白胆固醇（high density lipoprotein-cholesterol，HDL-C）水平的交互作用[22]。因此，作者针对不同发病率的疾病，假定其发生是基于一定的基因-环境交互作用，通过 QUANTO 软件[23]可以计算出所需的理论病例数，最终估算出不同样本量队列研究随访所需的最小年数，见表 2。（1）以发病率最高（约为 3.00/10 万人/年）的白内障、高血压为例，假定某基因型频率与某环境暴露率均为 0.15，为了识别效应值为 3.0 的基因-环境交互作用（乘法模型），理论上需要累计病例至少 1 173 例。此时，样本量为 20 万、50 万和 100 万的队列均可在 1 年内获取足够病例；（2）以发病率居中（约为 0.20/10 万人/年）的糖尿病、卒中、心力衰竭为例，其他条件相同，样本量为 20 万、50 万和 100 万的队列获取足够病例需分别随访 4 年、2 年和 1 年；（3）以发病率最小（约为 0.01/10 万人/年）的帕金森氏病，精神分裂症为例，相同条件下，100 万的队列需要随访 15 年，而样本量为 20 万和 50 万的队列即使随访 50 年也不能获取足够病例以发现基因 - 环境交互作用；（4）以上一种情况为例，其他条件不变，若改为识别效应值为 1.5 的基因-环境交互作用，则需要累积病例至少 8 743 例，此时即使是 100 万的队列在 50 年内也难以获取足够病例。因此，相比以往较为单一的基于环境因素或遗传因素的流行病学病因研究而言，为识别人类复杂性疾病的基因-环境交互作用，队列研究的样本量需达到数以万计、十万计，甚至是百万计的规模。

2 流行病学超大规模队列研究的发展

西班牙学者 Kogevinas[24]于 2002 年最早提出了整合欧洲出生队列的想法，并于 2004 年在其发表的一篇学术论文中最早使用了“Mega Cohort”一词[19]，呼吁将 7 个既存的以及规划中的欧洲出生队列整合成一个超过 50 万规模的超大出生队列。2005 年美国学者 Foster 和 Sharp[25]也认可流行病学超大规模队列研究在识别罕见遗传和环境因素对于复杂性疾病弱效应的优势。2010 年，JAMA 同一期刊登了 2 篇关于 Mega Cohort[26,27]的文章，从产生原因、研究设计、招募对象和结局确定等方面做了具体介绍。但至今为止，Mega Cohort 依然停留在字面意义上，样本量究竟多大才能被称为 Mega Cohort 并没有明确的规定，导致几万、几十万或上百万的队列均可称为 Mega Cohort。结合之前所介绍 Mega Cohort 的产生原因，本研究以 20 万人群样本量为最低标准列举了部分国内外 Mega Cohort，见表 3，并就其构建方式及优缺点进行一定的归纳和探讨。

表 1 前瞻性队列研究发病估计[18]

Table 1 Estimated disease incidence rates in prospective cohort studies[18]

发病数[率（/10 万人/年）]	疾病	不同样本量（S）队列在不同观察时间点（年）的估计发病数								
		S_1 = 5 000			S_2 = 50 000			S_3 = 500 000		
		5 年	10 年	20 年	5 年	10 年	20 年	5 年	10 年	20 年
10（0.01）	帕金森氏病、精神分裂症	2	4	7	23	42	74	228	424	737
50（0.05）	结肠直肠癌症、肾功能衰竭	11	21	37	114	212	367	1 141	2 118	3 672
100（0.10）	乳腺癌、髋部骨折	23	42	73	228	423	731	2 279	4 227	7 313
200（0.20）	糖尿病、卒中、心力衰竭	45	84	145	455	842	1 450	4 550	8 418	14 503
50（0.50）	心肌梗塞、癌症	113	208	354	1 131	2 078	3 535	11 309	20 780	35 354
3 000（3.00）	白内障、高血压	646	1 123	1 734	6 464	11 231	17 339	64 644	112 315	173 391

注：假定每年的损耗率（失访率）为 3%；来源于美国发病率患病率数据库。

表 2　前瞻性队列研究时间估计(以探索基因-环境交互作用的巢式病例对照研究设计研究)

Table 2　Time needed for matched pairs using "nested case-control" design in prospective cohort studies aiming for gene-environment interaction effects detection

基因-环境交互作用识别效应值	基因型频率	环境暴露率	所需的匹配病例对照数(对数)	不同样本量队列获得以下疾病所需病例数的最小年数(年)								
				帕金森氏病、精神分裂症(发病率=0.01/10 万人/年)			糖尿病、卒中、心力衰竭(发病率=0.20/10 万人/年)			白内障、高血压(发病率=3.00/10 万人/年)		
				20 万	50 万	100 万	20 万	50 万	100 万	20 万	50 万	100 万
1.5	0.05	0.05	44 389	–	–	–	–	–	41	10	3	2
	0.15	0.05	22 272	–	–	–	–	41	15	5	2	1
	0.45	0.05	28 007	–	–	–	–	–	20	6	2	1
	0.05	0.15	17 754	–	–	–	–	28	11	4	2	1
	0.15	0.15	8 743	–	–	–	40	11	5	2	1	1
	0.45	0.15	10 665	–	–	–	–	14	6	2	1	1
	0.05	0.45	11 184	–	–	–	–	15	7	2	1	1
	0.15	0.45	5 258	–	–	–	18	6	3	1	1	1
	0.45	0.45	5 870	–	–	–	21	7	4	2	1	1
3.0	0.05	0.05	5 157	–	–	–	18	6	3	1	1	1
	0.15	0.05	2 973	–	–	–	9	4	2	1	1	1
	0.45	0.05	4 699	–	–	–	16	6	3	1	1	1
	0.05	0.15	2 196	–	–	38	7	3	2	1	1	1
	0.15	0.15	1 173	–	–	15	4	2	1	1	1	1
	0.45	0.15	1 710	–	–	25	5	2	1	1	1	1
	0.05	0.45	1 642	–	–	24	5	2	1	1	1	1
	0.15	0.45	746	–	21	9	2	1	1	1	1	1
	0.45	0.45	829	–	24	10	3	1	1	1	1	1

注　"–"表示为了获得该类疾病所需的匹配病例对照数,在相应样本量队列中随访观察的时间 >50 年;表 2 的假定前提为:研究开始时队列中不存在任何患者,显性等位易感基因频率为 10%,环境暴露频率为 10%,年损耗率为 3%,把握度为 80%,Ⅰ类错误的概率为 0.000 1,遗传和环境边际效应为 1.5。

1976 年,美国的 Nurses' Health Study [28] 开启了当时针对女性职业人群全球最大规模的队列研究,虽然原始队列样本量不及 20 万,但第 2 队列的构建使得合计样本达到 238 387 例,而第 3 队列也于 2010 年启动,计划新招募 100 000 名美国或加拿大女护士。随着该项目的纵向发展,研究内容也由最初的口服避孕药对女性健康的影响逐渐延伸至当今人类复杂性疾病危险因素的研究领域 [28]。类似纵向发展的研究还有 Millennium Cohort Family Study (United States) [29],该研究于 2001 年、2004 年和 2007 年分别招募了第 1 批(77 047 例)、第 2 批(31 110 例)和第 3 批(43 440 例)研究对象,成为美国历史上最大的以军人为对象的前瞻性队列研究。此外,该研究又于 2011 年和 2012 年分别启动了第 4 批招募计划(约 60 000 例)以及 Millennium Cohort Family Study [29](约 10 000 例),进一步扩大研究样本和对象的范围,以待更全面的研究军人这一特殊职业人群的疾病和健康状况。在原始队列的基础上进行纵向发展,这种方式无疑是累积样本量以构建 Mega Cohort 经济可行且具有良好可持续性的手段,但需要注意的是由于不同纵向队列间研究对象的异质性及研究内容的保守性,会一定程度地限制数据的合并以及结论的外推。

1992 年,EPIC(The European Prospective Investigation into Cancer and Nutrition)研究 [30] 启动,至 2000 年共计纳入研究对象 521 468 例,涉及欧洲 10 个国家 23 个研究中心,作为欧盟抗癌规划(the Europe against cancer program of the european commission)的一部分,其研究内容已从起初营养与癌症的关系扩展到基因、心脏疾病、老年健康、空气污染等诸多领域,是全球少有的几个超过 50 万规模的 Mega Cohort 之一。基于欧洲各国的地域优势和联盟优势(如政治、经济以及科教文卫等领域),EPIC 研究将原本零散的、规模较小的和对象相对单一的队列研究整合成具有一定共性的 Mega Cohort,这种横向合并的方式在之后欧盟的诸多大型研究中被广泛应用。这种横向合并方式的最大优势在于能在较短的时间内形成多中心研究大样本,但需要注意的问题在于标化数据的获取和利用,是通过统一的方式前瞻性获取,还是通过查阅的方式回顾性整合,这

表 3 全球部分超大规模(样本量≥20 万)前瞻性流行病学队列研究列举

Table 3 A list of some prospective mega cohort studies at home and abroad with an minimum sample size of 200 000

研究名称	研究起始时间(年)	国家	队列入选对象特征	样本量
Nurses' Health Study				
Nurses' Health Study (original cohort) [15]	1976	美国	11 个州,年龄 30 ~ 55 岁的已婚注册女护士	121 701
Nurses' Health Study Ⅱ [15]	1989	美国	14 个洲,年龄 25 ~ 42 岁的女护士	116 686
Nurses' Health Study Ⅲ [28]	2010	美国,加拿大	2 个国家,年龄 20 ~ 46 岁的女护士或实习(学生)护士	100 000[a]
The European Prospective Investigation into Cancer and Nutrition(EPIC) [30,43]	1992	欧洲多国	10 个国家 23 个研究中心,年龄≥20 岁的居民	521 468
The NIH-AARP Diet and Health Study [31,44]	1995	美国	6 个洲和 2 个城市,年龄 50 ~ 69 岁的退休人员	567 169
The Million Women Study [32]	1996	英国	出生于 1932 - 1951 年,并于 1996 年 5 月 ~ 2001 年 3 月接受乳房 X 线照片检查的妇女	1 084 110
Millennium Cohort Study (United States) [29]				
Millennium Cohort Study (United States) [45,46]	2001	美国	全国所有军事机构的现役军人、预备军军人以及国民警卫队队员	211 597[a]
Millennium Cohort Family Study	2012	美国	于 2011 - 2012 年首次接受 Millennium Cohort Study 调研现役军人的(现任或曾经)伴侣	10 000[a]
The GenomEUtwin study[d] [47]	2002	多国	8 国双胞胎登记系统以及 MORGAM 队列研究中的对象	>600 000[c]
BioBank Japan Project for the Realization of Personalized Medicine [34]	2003	日本	57 家医院 47 种疾病中至少患有其中一种疾病的患者	200 000
The China Kadoorie Biobank [35]	2004	中国	5 个城市地区和 5 个农村地区,年龄 35 ~ 74 岁的常住居民	512 891
The New Generis Cohorts and Biobanks[d] [48]	2006	欧洲多国	大部分为产妇(来自 6 个国家的 8 个队列研究或生物银行项目)	300 000[b]
UK Biobank [36]	2007	英国	22 个社区中心,年龄 40 ~ 69 岁的志愿者	500 000
ENGAGE (European Network for Genetic and Genomic Epidemiology) [d] [49]	2008	欧洲多国	13 个国家,39 个队列研究中的对象	>600 000[b]
The National Cohort [50]	2012	德国	年龄 20 ~ 69 岁的德国人	200 000[a]
The Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) [51]	2012	美国	北加利福尼亚凯泽永久医疗集团(Kaiser Permanente)成员	500 000[a]
Million Veteran Program (MVP) [52]	2012	美国	美国退伍军人事务部保健系统的使用者	1 000 000[a]

注:[a] 进行或筹划中的样本量;[b] 估计样本量;[c] 估计的样本量对数;[d] 该研究已结束。

需要研究者综合多方面因素考虑。

1995 - 1996 年,美国国立卫生研究院(national institutes of health,NIH)和美国退休人员协会(formally the American association of retired persons,AARP)合作启动了 The NIH-AARP Diet and Health Study [31],该研究通过邮寄问卷(350 万份)的方式对美国 6 个洲及 2 个市 50 ~ 71 岁的 AARP 成员开展调查,最终形成了 567 169 例的 Mega Cohort。类似利用相关系统资源的研究还有英国的 The Million Women Study [32,33],基于英国国民医疗保健制度乳腺癌筛查项目(national health service breast cancer screening programme)相关内容——即每位 50 ~ 65 岁英国妇女每 3 年均会接收到该项目的邀请函,邀请她于之后的 2 ~ 6 周内到就近的乳腺癌筛查中心(the NHS breast-screening centre)进行 1 次乳房 X 线照片筛查,该研究于 1996 - 2001 年间通过搭载乳腺癌筛查邀请函的方式,共寄出约 200 万份调查问卷,最终纳入队列研究对象合计 1 084 110 例。通过利用大规模信息系统、组织部门、政策法规和相关活动等方式广泛募集研究对象,这种方式同样是构建 Mega Cohort 较为便利的方式之一。另外,作为其发展,如今利用相关卫生监测系统获取研究结局信息的方法,也被越来越广泛的应用在队列研究中。

作为开展个体化医疗和人类复杂性疾病病因研究(特别是基因-环境交互作用)的优质平台,这类基于生物样本库的 Mega Cohort 受到了生物医学界的广泛关注,如 BioBank Japan Project [34]、中国慢性病前瞻性研究项目 [35](China Kadoorie Biobank)和英国生物银行 [36](UK Biobank)。不仅如此,21 世纪 Mega Cohort 构建的方式也有了许多创新和发展,如:(1)队列募集的集中模式:相比于传统的分散模式(即长期设立各个项目地区研究中心),集中模式则是在各项目地区通过复制标准化招募过程(即通过预实验进行研究对象招募的演练,从而形成标准化的分工和流程 [37])以快速构建临时研究中心的方式,集中完成基线调查、数据传输和生物样本采集,随后便立即废除;(2)信息技术的综合利用:自行开发的 IT 系统,搭载触屏电脑进行无纸化问卷调查,可提高研究效率及信息准确性。传统队列研究是通过定期随访的方式以获取研究结局,但如今已逐渐被卫生监测系统(如死亡与恶性肿瘤登记、住院记录、初级卫生保健记录以及自报信息等 [38])的关联方式所取代;(3)重复调查:Mega Cohort 需在一定的时间间隔内从现存研究对象中抽取一定比例的样本开展重复调查,以评价暴露变化水平以及矫正回归稀释偏倚(regression dilution bias) [39]。还有诸多创

新的方法无法一一列举，但每种方法都是以服务于 Mega Cohort 的可行性和/或科学性为原则的。

就未来而言，Mega Cohort 必定需要源源不断的依赖于人力、物力、财力和时间的巨大投入和高新技术的不断发展[40]。以全基因组扫描(full genome sequencing)为例，美国 Illumina 公司 2011 年全基因组扫描的报价为 4 000 美元/人[41]，那么 50 万人仅此一项检测的费用就将花费 20 亿美元。所以，这类 Mega Cohort 目前均采取长期保存生物样本的方式以等待未来检测技术的提高及费用的降低，当纳入足够病例时再开展相关的遗传学研究。但何时条件成熟？生物样本再现性(reproducibility)[42]和变异性受保存期限的影响有多大？当前生物样本库的保存是否能满足未来生物医学技术的发展？未来流行病学领域新型研究方法的出现以及统计分析手段的进步是否会对 Mega Cohort 现有的信息提出新的要求？这一切既是机遇也是挑战。

3　总结

流行病学超大规模队列研究的构建方式各异，需基于特定的研究目的，因地制宜。每一种 Mega Cohort 的构建方式都有其自身的优缺点，不应盲目推崇某个最佳研究，而需弄清不同设计和方法背后的原因、条件和作用，从而指导研究者能更好的利用已有资源和/或创造必要条件以开展此类研究。Mega Cohort 并不是一种新兴的概念、研究或方法，也不是“把小研究做大”[26]的机械扩增，而是基于人类复杂性疾病病因研究的客观需求，在各方面条件(理论、实践、信息科技、生物医学技术、资金人力及其他可利用资源)都相对成熟的前提下，形成的一种开放友好且可持续发展的大规模综合平台(包括研究工具平台、生物资源平台和数据信息平台等)。Mega Cohort 的出现和发展不仅跨越了过去传统宏观研究与微观研究难以有机结合的鸿沟，更促进了学科交叉、技术融合和平台共享等先进理念的实践，更为重要的是它标志着生物医学领域大数据时代(the age of big data)[53]的到来。因此，需要利用流行病学超大规模队列研究以驱动发现和影响决策[54]，并期待它能带给 21 世纪生物医学带来革命性的突破[55]。

【参考文献】

[1] Ware J. Chirurgical Observations Relative to the Eye: With an Appendix, on the Introduction of the Male Catheter; and the Treatment of the H Morrhoids [M]. London: Printed for J. Mawma, 2010.

[2] ll'in MP. On the diagnosis of complex diseases of the heart [J]. Vrach Delo, 1967,8:39-44.

[3] Michael AF. Renal complex disease: streptococcus [J]. Zentralbl Bakteriol Orig, 1970,214(3):398-401.

[4] Eajans SS, Floyd JC Jr. Hypoglycemia: how to manage a complex disease [J]. Mod Med, 1973,41:24-31.

[5] Marx JL. Hypertension: a complex disease with complex causes [J]. Science, 1976,194(4267):821-825.

[6] Corrin B, Spencer H, Turner-Warwick M, et al. Pulmonary veno-occlusion-an immune complex disease? [J]. Virchows Arch A Pathol Anat Histol, 1974,364(1):81-91.

[7] Cello JP. Eosinophilic gastroenteritis: a complex disease entity [J]. Am J Med, 1979,67(6):1097-1104.

[8] Chakravarti A, Little P. Feature nature, nurture and human disease [J]. Nature, 2003:412-413.

[9] Lanktree MB, Hegele RA. Gene-gene and gene-environment interactions: new insights into the prevention, detection and management of coronary artery disease [J]. Genome Med, 2009,1(2):28.

[10] Risch N, Merikangas K. The future of genetic studies of complex human diseases [J]. Science, 1996,273(5281):1516-1517.

[11] Khoury MJ, Yang Q. The future of genetic studies of complex human diseases: an epidemiologic perspective [J]. Epidemiology, 1998,9(3):350-354.

[12] Raymond CA. Third World's Complex Disease Problems Compounded by Economic, Cultural Factors [J]. JAMA, 1988,260(24):3557-3561.

[13] Newman B, Lee M, Stillman L, et al. Gene mapping of a simulated complex disease [J]. Prog Clin Res, 1989,329:171-176.

[14] World Health Organization. Global status report on noncommunicable diseases 2010 [R]. Geneva, 2011.

[15] Colditz GA, Manson JE, Hankinson SE. The Nurses' Health Study: 20-year contribution to the understanding of health among women [J]. Journal of Women's Health, 1997,6(1):49-62.

[16] Tamakoshi A, Yoshimura T, Inaba Y, et al. Profile of the JACC study [J]. J Epidemiol, 2005,15(Suppl 1):S4-S8.

[17] International Agency For Research On Cancer W, Europe Against Cancer Commission E. EPIC-European Prospective Investigation into Cancer and Nutrition [K]. 2012.

[18] Manolio TA, Bailey-Wilson JE, Fraricis S. Collins. Genes, environment and the value of prospective cohort studies [J]. Nature Reviews Genetics, 2006,7(10):812-820.

[19] Kogevinas M, Andersen AM, Olsen J. Collaboration is needed to co-ordinate European birth cohort studies [J]. Int J Epidemiol, 2004,33(6):1172-1173.

[20] 叶冬青. 巢式病例对照研究的设计及分析 [J]. 疾病控制杂志, 2001,5(1):65-68.

[21] Vasan RS, Beiser A, Seshadri S, et al. Residual lifetime risk for developing hypertension in middle-aged women and men [J]. JAMA, 2002,287(8):1003-1010.

[22] Ordovas JM, Corella D, Demissie S, et al. Dietary fat intake determines the effect of a common polymorphism in the hepatic lipase

gene promoter on high-density lipoprotein metabolism: evidence of a strong dose effect in this gene-nutrient interaction in the Framingham Study [J]. Circulation, 2002,106(18):2315-2321.

[23] Gauderman WJ. Sample size requirements for matched case-control studies of gene - environment interaction [J]. Stat Med, 2002, 21(1):35-50.

[24] Kojevinas M. Expression of interest for an integrated project: European Union birth-cohort on child health and human development. Luxembourg: Community Research and Development Information Service [K]. 2002.

[25] Foster MW, Sharp RR. Will investments in large-scale prospective cohorts and biobanks limit our ability to discover weaker, less common genetic and environmental contributors to complex diseases? [J]. Environ Health Perspect, 2005,113(2):119-122.

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

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

[28] Speizer F. Nurses' Health Study [K]. 2012.

[29] Millennium Cohort Family Study. The Millennium Cohort Family Study is a Department of Defense research project at the Deployment Health Research Department [K]. 2012.

[30] 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.

[31] Schatzkin A, Subar AF, Thompson FE, et al. Design and serendipity in establishing a large cohort with wide dietary intake distributions: the National Institutes of Health-American Association of Retired Persons Diet and Health Study [J]. Am J Epidemiol, 2001,154(12):1119-1125.

[32] Whitehead M, Farmer R. The million women study: a critique [J]. Endocrine, 2004,24(3):187-193.

[33] Beral V, Million Women Study Collaborators. Breast cancer and hormone-replacement therapy in the Million Women Study [J]. Lancet, 2003,362(9392):419-427.

[34] Nakamura Y. BioBank Japan Progect for the Realization of Personalized Medicine [K]. 2012.

[35] 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.

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

[37] UK Biobank Coordinating Centre. UK Biobank: Report of the integrated pilot phase [K]. 2006.

[38] UK Biobank Coordinating Centre. UK Biobank: Protocol for a large-scale prospective epidemiological resource [K]. 2007.

[39] Manolio TA, Weis BK, Cowie CC, et al. New Models for Large Prospective Studies: Is There a Better Way? [J]. Am J Epidemiol, 2012,176(11):1-8.

[40] Collins FS. The case for a US prospective cohort study of genes and environment [J]. Nature, 2004,429(6990):475-477.

[41] Illumina. Illumina Announces $ 5,000 Genome Pricing [K]. 2011.

[42] Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine [J]. Int J Epidemiol, 2008,37(2):234-244.

[43] Gonzalez CA. The European prospective investigation into cancer and nutrition (EPIC) [J]. Public health nutrition, 2006,9(1A):124-126.

[44] Park Y, Brinton LA, Subar AF, et al. Dietary fiber intake and risk of breast cancer in postmenopausal women: the National Institutes of Health - AARP Diet and Health Study [J]. The American journal of clinical nutrition, 2009,90(3):664-671.

[45] Leleu TD, Jacobson IG, Leardmann CA, et al. Application of latent semantic analysis for open-ended responses in a large, epidemiologic study [J]. BMC medical research methodology, 2011, 11(1):136.

[46] Smith TC. Linking Exposures and Health Outcomes to a Large Population-Based Longitudinal Study: The Millennium Cohort Study [J]. Military medicine, 2011,176(7 Suppl):56-63.

[47] Peltonen L. GenomEUtwin: a strategy to identify genetic influences on health and disease [J]. Twin Res, 2003,6(5):354-360.

[48] Merlo DF, Wild CP, Kogevinas M, et al. NewGeneris: a European study on maternal diet during pregnancy and child health [J]. Cancer Epidemiology Biomarkers Pre, 2009,18(1):5-10.

[49] Tassé AM, Budin-Ljøsne I, Knoppers BM, et al. Retrospective access to data: the ENGAGE consent experience [J]. Eur J Hum Genet, 2010,18(7):741-745.

[50] A network of German research institutes from the Helmholtz Association, the Leibniz Association. The National Cohort [K]. 2012.

[51] Schaefer C. The Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) [K]. 2012.

[52] Michael E. DeBakey VA Medical Center. Million Veteran Program: A Partnership with Veterans [K]. 2012.

[53] Lohr S. The age of big data [EB/OL]. (2012-02-12) [2012-10-01]. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html? src = mv&ref = general [/url].

[54] World Economic Forum. Big Data, Big Impact: New Possibilities for International Development [R]. Geneva, 2012.

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

(收稿日期:2012-08-13)
(修回日期:2012-11-27)
(刘岩岩校)