American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2012.

Vol. 175, No. 9
DOI: 10.1093/aje/kwr453
Advance Access publication:
March 12, 2012

Commentary

New Models for Large Prospective Studies: Is There a Better Way?

Teri A. Manolio*, Brenda K. Weis, Catherine C. Cowie, Robert N. Hoover, Kathy Hudson, Barnett S. Kramer, Chris Berg, Rory Collins, Wendy Ewart, J. Michael Gaziano, Steven Hirschfeld, Pamela M. Marcus, Daniel Masys, Catherine A. McCarty, John McLaughlin, Alpa V. Patel, Tim Peakman, Nancy L. Pedersen, Catherine Schaefer, Joan A. Scott, Timothy Sprosen, Mark Walport, and Francis S. Collins

* Correspondence to Dr. Teri A. Manolio, Office of Population Genomics, National Human Genome Research Institute, 5635 Fishers Lane, Suite 3058, MSC 9307, Bethesda, MD 20892-9307 (e-mail: manolio@nih.gov).

Initially submitted May 23, 2011; accepted for publication November 9, 2011.

Large prospective cohort studies are critical for identifying etiologic factors for disease, but they require substantial long-term research investment. Such studies can be conducted as multisite consortia of academic medical centers, combinations of smaller ongoing studies, or a single large site such as a dominant regional health-care provider. Still another strategy relies upon centralized conduct of most or all aspects, recruiting through multiple temporary assessment centers. This is the approach used by a large-scale national resource in the United Kingdom known as the ''UK Biobank,'' which completed recruitment/examination of 503,000 participants between 2007 and 2010 within budget and ahead of schedule. A key lesson from UK Biobank and similar studies is that large studies are not simply small studies made large but, rather, require fundamentally different approaches in which ''process'' expertise is as important as scientific rigor. Embedding recruitment in a structure that facilitates outcome determination, utilizing comprehensive and flexible information technology, automating biospecimen processing, ensuring broad consent, and establishing essentially autonomous leadership with appropriate oversight are all critical to success. Whether and how these approaches may be transportable to the United States remain to be explored, but their success in studies such as UK Biobank makes a compelling case for such explorations to begin.

cohort studies; epidemiology; prospective studies

Large prospective cohort studies are indispensable for identifying etiologic factors for disease, but their large scope requires substantial long-term research investments. Prospective cohort studies are often conducted as multisite consortia of large academic medical centers, each responsible for recruitment, examination, and follow-up of participants in its geographic area. This was the model initially explored for a prospective study of up to 1,000,000 Americans (1), but its large anticipated costs present a substantial hurdle.

Other models include combining ongoing, smaller studies to provide more immediate answers at potentially lower costs (2, 3), but existing studies are often limited and variable in the diversity of populations, exposures, and diseases ascertained; standardization of methods; and adequacy of existing consent and data access (4, 5). Large cohorts recruited at a single site, often through a dominant regional health-care provider, represent another approach (6, 7). Still another strategy relies upon centralized conduct of most or all aspects of the study, although responsibility for individual study-wide aspects can be distributed to collaborating centers (Figure 1) (8). This is the approach used by a large-scale national resource in the United Kingdom known as ''UK Biobank'' (9), initiated in 2004 to examine genetic and environmental risk factors for complex diseases. Briefly, UK Biobank issued mailed invitations to about 9 million persons aged 40–69 years living within 25 miles (40.23 km) from the 21 study assessment centers and registered with the National Health Service in the United Kingdom. More than 500,000 participants responded to this invitation and were examined between April 2007 and June 2010. Detailed questionnaire, interview, and measurement data were obtained during the 2–3 hour baseline assessment, and multiple samples of blood, saliva, and urine were collected (Table 1). Repeat assessments on a subsample of 20,000 participants will address regression-dilution bias,

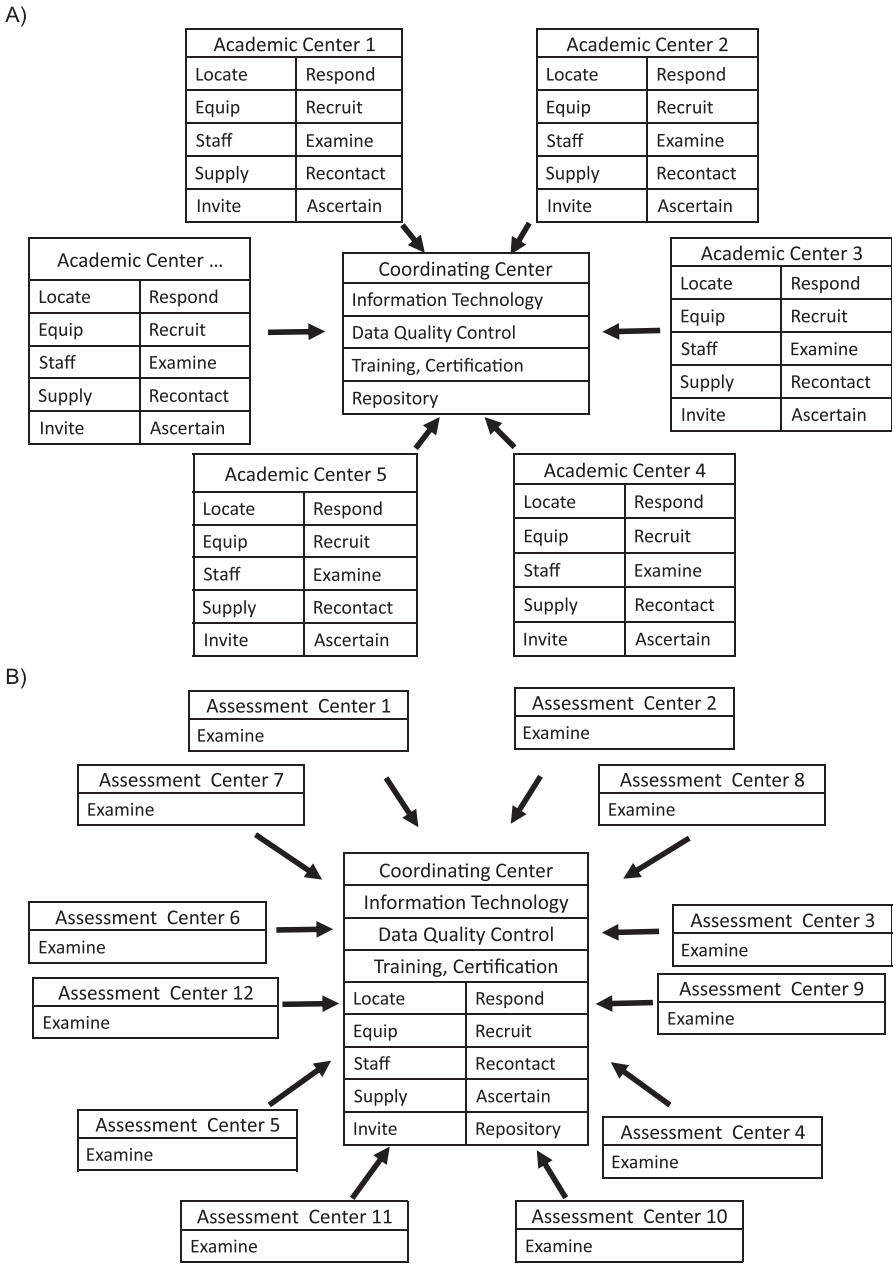

Figure 1. Traditional distributed model contrasted with novel centralized model. In A (a traditional distributed model), the roles of collaborating, typically academic, centers in distributed models include large numbers of tasks that could be located in one or more central units. Coordinating centers in these models tend to be responsible for developing and implementing data collection and monitoring systems, training and certification standards for study staff, and the biospecimen repositories. In B (a novel centralized model), centralized models concentrate multiple study-wide activities in one or more coordinating centers, with each potentially handling related clusters of activities. Assessment centers focus on participant examination and transmission of data and specimens, and they are kept open only so long as they are productive. Successive waves of assessment centers represented by an initial group of centers 1–6 are followed by a wave of centers 7–12, and so on.

and health outcomes will be assessed primarily through national health records.

The centralized model, which UK Biobank adopted after rejecting a costly, decentralized approach, has enabled its leaders to achieve exceptional efficiencies in recruitment, assessment, and record linkage while retaining diversity in demographics and exposures. Cohort recruitment was achieved not only ahead of schedule but also within budget (roughly $100 million for 503,000 participants), enabling substantial enhancements to be added to the study protocol.

The evident success of the UK Biobank model may provide valuable lessons for the conduct of large-scale studies

Table 1. Components of UK Biobank, With Participant Recruitment between 2007 and 2010

Baseline Questionnaire	Baseline Physical Measurements	Follow-up and Future Measures
Sociodemographic	Blood pressure	Stored blood, urine, saliva
Family history	Weight, body impedance	Repeat baseline assessment (20,000 participants)
Psychosocial	Waist and hip circumferences	Access national health records
Environmental	Seated and standing heights	• Death
Lifestyle	Grip strength	• Cancer
Cognitive function	Spirometry	• Hospitalizations
Health status	Bone density	• Primary care
Food frequency	Mailed triaxial accelerometers	
Internet-administered 24-hour dietary questionnaire	Enhanced phenotyping (last 100,000–150,000 participants recruited)	
	• Hearing	
	• Vascular reactivity	
	• Visual acuity	
	• Refractive error	
	• Intraocular pressure	
	• Corneal biomechanics	
	• Optical coherence tomography	
	• Fitness assessment	

Abbreviation: UK, United Kingdom.

that may or may not be feasible to implement in the United States. The National Institutes of Health thus convened a symposium summarized in this paper, involving several large studies (Table 2) (10), to 1) examine novel aspects of the UK Biobank design and its strengths and weaknesses; 2) compare the UK Biobank approach with those of other large studies; and 3) identify lessons learned in maximizing the efficiencies of these studies.

OVERARCHING CONSIDERATIONS

Suggested characteristics of an optimal cohort study for examining genetic and environmental influences on disease have been described (Table 3) (1). Large size is a key component, as relevant genetic variants and other risk exposures may be uncommon and effect sizes are often modest (11). These are not, however, simply small studies made large, as the costs and inefficiencies in 100-fold expansion of a 5,000-person, disease-specific cohort study are prohibitive. Large studies require fundamentally different approaches in which minimizing cost is a primary consideration and ''process'' expertise to maximize efficiency of high-throughput operations is as important as scientific rigor. Modern industrial design principles that identify and manage critical choke points are essential to ensuring high throughput and maintaining quality (12).

Decentralized models involving semipermanent research centers can be expensive to maintain and can present challenges in standardization. Using temporary assessment centers in a centralized model avoids the need to maintain remote offices, staff, and laboratory capabilities. Centralized models may provide greater overall control of costs, as well as agility in responding to changing situations such as relocating underperforming sites or modifying suboptimal procedures. They may thus free investigators to focus on science rather than miring them in the operational concerns of their individual sites.

Centralized models may also have drawbacks. The inherent need to choose a specific population base and standardized assessments for a very large, centralized cohort may limit the questions that can be addressed. This contrasts with the diverse approaches fostered by multiple independent studies that can be a powerful force for improving methodology and assessing the replicability of findings. In addition, the potential for disenfranchising academic centers accustomed to operational leadership in their assigned geographic area may risk losing critical scientific input from these groups. Special care is also needed to involve community-based organizations and to ensure that they feel local concerns are being addressed in a centralized study design. Keeping them engaged in a centralized model requires significant effort with frequent local visits and community meetings that include the study leadership. Being chosen as part of a major national effort can be a source of considerable community pride, especially if it is clear that community input is valued and implemented.

Experienced investigators may also have well-functioning local recruitment systems and understanding of unique local conditions that may require tailoring of methods. Approaches for harnessing this expertise need careful attention, but they might include engaging academic investigators in protocol development and implementation or tasking individual academic centers with study-wide functions such as ensuring diversity of participants, developing novel substudies, or responding to queries through a participant call center.

Table 2. Large Studies Examined

Decentralized, Multisite	Single Site	Centrally Coordinated
Canadian Partnership for Tomorrow	Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH)	American Cancer Society Cancer Prevention Study 3 (ACS CPS-3)
European Prospective Investigation into Cancer and Nutrition (EPIC)	Marshfield Clinic Personalized Medicine Research Program (PMRP)	LifeGene
National Children's Study	Vanderbilt BioVU, a research resource providing a "view into biology" at the level of DNA	National Health and Nutrition Examination Survey (NHANES)
Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial		UK Biobank
VA Genomic Medicine Program		
Women's Health Initiative		

Abbreviation: UK, United Kingdom.

A key aspect of limiting costs is the vigor with which a high "response rate" is pursued during recruitment. Nationally representative surveys, such as the National Health and Nutrition Examination Survey (NHANES) (13) and smaller disease-specific studies such as the Cardiovascular Health Study (14), serve important aims in providing population-based estimates of disease prevalence and incidence. For this purpose, they need representative population samples, necessitating considerable expenditures to ensure high response rates (15, 16). The limitations of an essentially volunteer sample are well known, particularly the generally healthier profile, higher educational attainment, and greater health consciousness of volunteers (17). This "healthy volunteer" effect can lead to underestimation of disease prevalence and incidence, but its impact on relative risk estimates for environmental and genetic factors is generally not important (18, 19). Although high response rates are critical for population-based estimates of disease incidence, prevalence, or mortality, a 10% or even 1% response rate may be acceptable in certain situations, especially if the focus is on risk associations and the base population is large enough to capture a diversity of exposures and backgrounds. Results from such studies can still be applicable to populations with different distributions of these exposures, although this cannot be proven but only assumed for exposures that are unknown or unknowable. High response rates thus need not be a driving factor scientifically or economically. For these reasons, UK Biobank chose to emphasize diversity but to de-emphasize response rates. It has accepted yields of 5%–10% while realizing substantial savings by not attempting to convert initial refusals, as conversion was not found to be effective in pilot studies.

Other important cost determinants include method of ascertainment (such as registry vs. household enumeration), complexity of data collection, and follow-up methods (such as active vs. passive). Choice among these is largely driven by the scope and goals of a given study. Irrespective of these choices, however, centralized approaches are likely to provide the advantages described above, as evidenced by the marked reductions in UK Biobank costs when it shifted to a centralized design while keeping other aspects constant.

Table 3. Characteristics of Optimal Cohort Study[a]

- Large in scale (hundreds of thousands of participants)
- Diverse regarding age, race/ethnicity, socioeconomic status, geographic region
- Address multiple diseases/risk factors
- Highly efficient recruitment, data collection, sample processing
- Standardized or harmonized terminology to facilitate interoperability with other data
- Linked personal electronic records and biospecimens
- Broad content and high quality of samples and data
- State-of-the-art technology for environmental sampling, laboratory methods, genomics, information technology
- Cost effective
- Data available for qualified researchers

[a] According to F. S. Collins (*Nature*. 2004;429(6990):475–477) (1).

RECRUITMENT

Embedding participant recruitment within a structure that facilitates determination of study outcomes, such as an existing electronic medical records system (20) or a reimbursement system (16), greatly simplifies participant follow-up. The Kaiser Research Program on Genes, Environment, and Health (RPGEH), Marshfield Clinical Personalized Medicine Research Program (PMRP), UK Biobank, and Vanderbilt University BioVU all recruit from their respective patient registries, simplifying subsequent record linkage for follow-up. As it can be challenging to use such information to address research hypotheses, such data may need further review and validation.

Centralizing the invitation and appointment system can maximize efficiencies in recruitment and scheduling, allowing study-wide monitoring of invitations sent, confirmations received, and appointments made. A centralized approach also facilitates weekly or even daily monitoring of key characteristics of participants study-wide; sampling fractions can then be modified in real time to increase recruitment of undersubscribed groups.

Recognizing that participation rates tend to fall with increased distance to an examination center, invitations can be concentrated close to the center and extended gradually

outward until participation falls enough to justify closing that center and opening another in a new area (Web Figure 1, posted on the *Journal*'s website (http://aje.oxfordjournals.org/)). High-performing centers can also be kept open longer than originally planned as long as they remain cost-effective and provide sufficient diversity in exposures and demographics. Optimal locations for assessment centers can be pinpointed by mapping population density, transportation routes, and available office space. Establishing assessment centers where and when they are needed, staffing them primarily with temporary personnel specifically trained for the needs of the project, and closing them when they cease to meet recruitment goals provide considerable efficiencies over establishing a fixed number of centers and keeping all of them open throughout the recruitment period. Cost per participant recruited is a driving force in the cost of these studies and should be a primary focus in designing recruitment strategies, as long as fundamental scientific goals remain paramount.

Novel approaches to recruitment include enrolling participants at fund-raising events, such as the American Cancer Society's ''Relay For Life'' or through worksite recruitment at corporate partners of such organizations. Yield of these strategies can be low, however, particularly for certain demographic groups. Direct mail approaches may also introduce some demographic biases because willingness to respond to, or indeed even to open, mailed invitations varies considerably. Even a 1% yield, however, as experienced in some US studies (21), may be cost-effective if mailing is inexpensive and there is a sufficiently large and diverse base population. Internet and other social media approaches have potential for even lower costs and are being explored in many studies, but they do have pronounced differences by demographics; this may abate over time as Internet use expands. Text messaging for confirmation of appointments, for example, has proven to be effective with younger participants (22).

Another aspect that can be usefully centralized is responding to queries about the study. A single participant call center can rapidly gain experience in responding to concerns effectively and consistently, and it can also allow rapid escalation to more senior staff as needed. In UK Biobank, for example, roughly 50% of invitees calling to express concerns ultimately participated, and data on specific concerns were used to modify invitation materials and clinic flow (23).

Sending invitees a provisional appointment (''. . . an appointment has been scheduled for you on date/time; if you would like to confirm or change this optional appointment please contact us . . .'') entails some risk, as they may view it as presumptuous. It has, however, been used successfully by some health services for their national screening programs, such as National Health Service breast screening in the United Kingdom. Like all other aspects of recruitment, it needs to be carefully piloted, and its effectiveness may well vary by cultural or demographic characteristics. Still, it was quite successful in UK Biobank where roughly half of all participants who attended did so at the time of their provisional appointments.

QUESTIONNAIRE AND EXAMINATION

Several studies, including the European Prospective Investigation into Cancer and Nutrition (EPIC), the US National Health and Nutrition Examination Survey, and UK Biobank, have recognized the efficiency of having participants respond to questionnaires by touch screen rather than through interviewer- or self-administered forms. Well-designed touchscreen systems can allow a single staff member to assist 15–20 participants at a time. Electronic data capture in real time minimizes errors in transcription, and ongoing analysis of data quality can help to identify and correct problems early. In addition, computerized instruments can be used to collect more than questionnaire data, including tests of cognitive function, audiology, and simple motor function. Internet-based 24-hour dietary recalls administered remotely in UK Biobank, for example, will permit more in-depth evaluation of intake of total energy and some nutrients and will facilitate repeat measurements.

BIOREPOSITORY

Biorepositories provide much of the future value of cohort studies, as they permit measurement of biomarkers of exposures or intermediate phenotypes often not even imagined at the time the study is initiated. Central processing of biospecimens typically offers increased consistency and achievable throughput, a robust data trail, and lower costs compared with local processing, although this must be piloted carefully in each setting (12). Automated, industrial-scale specimen processing and storage systems are indispensable in large studies although their costs may make them impractical for smaller studies. Implementation following best practices used in the manufacturing industry can reduce project risk and build in quality and robustness (12).

HEALTH OUTCOMES

Large-scale studies can examine a nearly unlimited number of outcomes provided the outcomes can be classified reliably

Table 4. Key Lessons From New Models of Large Cohort Studies

- Ensure that future studies, including disease-specific studies, address the widest possible range of outcomes to permit combining data for increased study power
- Use standardized or harmonized (not identical but comparable) measures to permit diverse studies to be combined[a]
- Establish consents that allow for broad data sharing as the norm
- Maximize cost-efficiency where appropriate by
  - Exploring centralized recruitment and examination models
  - Considering lower recruitment yield if associations rather than prevalence are the primary objective
  - Utilizing electronic records
  - Emphasizing industrial-scale process expertise as the driver of process organization, implementation, and monitoring
  - Maximizing the capabilities of information technology to ensure high-quality data, rapid transfer, and real-time monitoring
  - Phasing activities to be completed only shortly before they are needed

[a] According to DataSHaPER (http://www.datashaper.org/) (30) and C. M. Hamilton et al. (doi:10.1093/aje/kwr193) (31).

and sufficient numbers of cases are available. Collection of outcome data is greatly facilitated by comprehensive electronic medical records; indeed, the feasibility of UK Biobank was questioned before links with National Health Service outcomes data were firmly established. Electronic medical records can pose unique challenges, however, because they often include complex and variable medical terminology, may be difficult to interpret with regard to timing of measurements, and cannot easily be reduced to common elements for pooled analyses.

To the degree that standardized terminology can be extracted from electronic medical records, however, automated algorithms can be developed and are being actively used in programs such as the Marshfield and Vanderbilt studies (24, 25). Extending electronic systems of outcome assessment to a national scale may be one of the greatest challenges to implementing a truly nationwide US cohort study.

INFORMATION TECHNOLOGY

High-quality information technology systems arguably represent the single most important infrastructure component for ensuring high data quality and cost control in large studies. Ideally, they should be designed to maximize participant throughput, perform routine data quality checks, and enable access and linkage to other data systems.

Real-time data capture and essentially daily review are critical to the smooth operations of large-scale projects, as the volume of data and potential errors identified through less frequent monitoring could become overwhelming. Use of commercial software when possible may be cost-effective, particularly for routine components such as laboratory information systems, but such systems must be interoperable with other assessment center systems. Working within established electronic medical record systems can facilitate leveraging extensive infrastructures established for clinical care. Research use of these systems may also facilitate further development of electronic medical records for research and, potentially, for direct incorporation of research findings into clinical care.

Threats to participant privacy may best be managed by removing obvious identifiers and establishing user agreements cosigned by an outside data user's institution, with reliance on professional standards of conduct to prevent misuse (26). Minimizing reidentification risk is an area of active investigation (27).

CONSENT AND COMMUNITY CONCERNS

Maximizing the value of large-scale studies requires ready access by a wide community of investigators for research consistent with participants' consent. Studies not requiring a high recruitment yield can include broad data collection, sharing, and use in their initial consent process and exclude invitees uncomfortable with these terms. Other studies may allow multiple levels of consent designating permitted research uses and data acquisition, and use can then be customized to the level of consent. Funders of these studies also have a responsibility, and a clear interest, in maximizing the value of the research investment by ensuring broad data sharing within the constraints of participant privacy and consent. Participants' expectations for return of research results may also need to be addressed (28).

External oversight helps to ensure that participants' concerns are voiced and addressed. Many such studies have community advisory boards that meet regularly, while larger studies may have more formally chartered and even nationally recognized advisory groups such as UK Biobank's independent Ethics and Governance Council. Such groups essentially ''hold up a mirror'' to ongoing studies and provide valuable input as well as wider credibility on issues such as privacy, adherence to consent limitations, and ethics. Funders should maintain some distance from ethics and governance issues, yet cannot remain totally aloof, as large studies generally carry the imprimatur of the funding organization and potential lapses will often be laid at their door.

CONCLUSIONS AND POTENTIAL NEXT STEPS

That large prospective studies are not simply small studies made large is a crucial consideration in successfully carrying them out. The scale of these studies necessitates entirely different approaches, emphasizing cost-efficiency to a degree that may appear inimical to the scientific method. In fact, it can be indispensable to the science, enabling studies to go forward that might otherwise be impossible. The £68 million (~$100 million) cost of developing and recruiting the UK Biobank cohort over 3 years stands in stark contrast to the roughly $400 million per year for 10 years estimated for traditional models of conducting a large cohort study in the United States (29).

Cohort studies, whether large or small, have much to learn from and contribute to each other (Table 4). As all models have their strengths and weaknesses, it is unwise to propose from among them the one ''best,'' but better to recognize the need for, and complementary nature of, a variety of designs and approaches. Adoption of a large, centralized model for one study should not, by virtue of its size and cost, have a chilling effect on enthusiasm for and funding of other complementary approaches.

Methods successfully used in European studies, such as the European Prospective Investigation into Cancer and Nutrition, LifeGene, and UK Biobank, may not all be directly transportable to the United States, where infrastructures for facilitating follow-up on a national scale are limited and where diversity in lifestyle, ancestral origin, and geography, as well as sheer size, is considerably greater than in many other countries. Careful piloting and critical evaluation will be needed to assess the feasibility of such models in the United States. The success to date of these models in studies such as UK Biobank, however, makes a compelling case for such explorations to begin.

ACKNOWLEDGMENTS

Author affiliations: Office of Population Genomics, National Human Genome Research Institute, Bethesda, Maryland (Teri A. Manolio); Division of Program Coordination, Planning, and Strategic Initiatives, National Institutes of

Health, Bethesda, Maryland (Brenda K. Weis, Barnett S. Kramer); National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, Maryland (Catherine C. Cowie); Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland (Robert N. Hoover); Office of the Director, National Institutes of Health, Bethesda, Maryland (Kathy Hudson, Francis S. Collins); Division of Cancer Prevention, National Cancer Institute, Bethesda, Maryland (Chris Berg); Clinical Trial Service Unit and Epidemiological Studies Unit, University of Oxford, Oxford, United Kingdom (Rory Collins); Medical Research Council, London, United Kingdom (Wendy Ewart); Boston VA Medical Center, Boston, Massachusetts (J. Michael Gaziano); National Institute of Child Health and Human Development, Bethesda, Maryland (Steven Hirschfeld); Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland (Pamela M. Marcus); Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee (Daniel Masys); Essentia Institute for Health Research, Duluth, Minnesota (Catherine A. McCarty); Cancer Care Ontario and the Ontario Institute for Cancer Research, Toronto, Ontario, Canada (John McLaughlin); Cancer Prevention Study 3, American Cancer Society, Atlanta, Georgia (Alpa V. Patel); UK Biobank, Cheshire, United Kingdom (Tim Peakman); Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden (Nancy L. Pedersen); Kaiser Permanente Program on Genes, Environment, and Health, Oakland, California (Catherine Schaefer); National Coalition for Health Professional Education in Genetics, Washington, DC (Joan A. Scott); Imperial College, London, United Kingdom (Timothy Sprosen); and Wellcome Trust, London, United Kingdom (Mark Walport).

This paper summarizes the deliberations of a symposium convened and supported by the National Institutes of Health on January 22, 2010, to examine new models for conducting large-scale prospective cohort studies.

Conflict of interest: none declared.

REFERENCES

1. Collins FS. The case for a US prospective cohort study of genes and environment. *Nature*. 2004;429(6990):475–477.
2. Hunter DJ, Riboli E, Haiman CA, et al. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. National Cancer Institute Breast and Prostate Cancer Cohort Consortium. *Nat Rev Cancer.* 2005;5(12):977–985.
3. Riboli E, Kaaks R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol*. 1997;26(suppl 1):S6–S14.
4. Willett WC, Blot WJ, Colditz GA, et al. Merging and emerging cohorts: not worth the wait. *Nature*. 2007;445(7125):257–258.
5. Collins FS, Manolio TA. Merging and emerging cohorts: necessary but not sufficient. *Nature*. 2007;445(7125):259. (doi:10.1038/445259a).
6. McCarty CA, Wilke RA, Giampietro PF, et al. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. *Per Med*. 2005;2(1):49–79.
7. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther.* 2008;84(3):362–369.
8. Manolio TA, Collins R. Enhancing the feasibility of large cohort studies. *JAMA*. 2010;304(20):2290–2291.
9. biobank^{uk}. Stockport, England: UK Biobank; 2011. (http://www.ukbiobank.ac.uk/). (Accessed September 18, 2011).
10. NIH Common Fund. Large-scale cohort studies. Bethesda, MD: National Institutes of Health; 2012. (http://commonfund.nih.gov/newmodels/meetings/newmodels012210/index.aspx). (Accessed September 18, 2011).
11. Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. *Nat Rev Genet.* 2006;7(10):812–820.
12. Downey P, Peakman TC. Design and implementation of a high-throughput biological sample processing facility using modern manufacturing principles. *Int J Epidemiol*. 2008;37(suppl 1):i46–i50.
13. Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey. Hyattsville, MD: National Center for Health Statistics; 2011. (http://www.cdc.gov/nchs/nhanes.htm). (Accessed September 18, 2011).
14. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. *Ann Epidemiol*. 1991;1(3):263–276.
15. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC Investigators. *Am J Epidemiol.* 1989;129(4):687–702.
16. Tell GS, Fried LP, Hermanson B, et al. Recruitment of adults 65 years and older as participants in the Cardiovascular Health Study. *Ann Epidemiol.* 1993;3(4):358–366.
17. Holden G, Rosenberg G, Barker K, et al. The recruitment of research participants: a review. *Soc Work Health Care*. 1993;19(2):1–44.
18. Stolley PD, Schlesselman JJ. *Case-Control Studies: Design, Conduct, and Analysis*. Oxford, United Kingdom: Oxford University Press; 1982:128–129.
19. Morton LM, Cahill J, Hartge P. Reporting participation in epidemiologic studies: a survey of practice. *Am J Epidemiol.* 2006;163(3):197–203.
20. Kaiser Permanente. The Research Program on Genes, Environment, and Health (RPGEH). Oakland, CA: Kaiser Permanente Division of Research; 2011. (http://www.dor.kaiser.org/external/DORExternal/rpgeh/index.aspx). (Accessed September 18, 2011).
21. Gren L, Broski K, Childs J, et al. Recruitment methods employed in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. *Clin Trials*. 2009;6(1):52–59.
22. LifeGene Sweden. Stockholm, Sweden: Karolinska Institutet; 2011. (https://www.lifegene.se/In-english/). (Accessed September 18, 2011).
23. UK Biobank. Report of the integrated pilot phase. Cheshire, United Kingdom: UK Biobank Coordinating Centre; 2006. (http://www.ukbiobank.ac.uk/docs/IntegratedPilotReport.pdf). (Accessed September 18, 2011).
24. Wilke RA, Berg RL, Linneman JG, et al. Quantification of the clinical modifiers impacting high-density lipoprotein cholesterol in the community: Personalized Medicine Research Project. *Prev Cardiol*. 2010;13(2):63–68.
25. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26(9):1205–1210.
26. Manolio TA, Rodriguez LL, Brooks L, et al. New models of collaboration in genome-wide association studies: the Genetic

Association Information Network. GAIN Collaborative Research Group; Collaborative Association Study of Psoriasis; International Multi-Center ADHD Genetics Project; Molecular Genetics of Schizophrenia Collaboration; Bipolar Genome Study; Major Depression Stage 1 Genomewide Association in Population-based Samples Study; Genetics of Kidneys in Diabetes (GoKinD) Study. *Nat Genet*. 2007;39(9):1045–1051.
27. Loukides G, Gkoulalas-Divanis A, Malin B. Anonymization of electronic medical records for validating genome-wide association studies. *Proc Natl Acad Sci U S A*. 2010;107(17):7898–7903.
28. Murphy J, Scott J, Kaufman D, et al. Public expectations for return of results from large-cohort genetic research. *Am J Bioeth.* 2008;8(11):36–43.
29. Design considerations for a potential United States population-based cohort to determine the relationships among genes, environment, and health: recommendations of an expert panel. Bethesda, MD: National Human Genome Research Institute; 2005. (http://www.genome.gov/Pages/About/OD/ReportsPublications/PotentialUSCohort.pdf). (Accessed September 18, 2011).
30. Data Schema and Harmonization Platform for Epidemiological Research (DataSHaPER). Montreal, Canada: P^3G Observatory, The Research Institute of the McGill Health Center; 2011. (http://www.datashaper.org/). (Accessed September 18, 2011).
31. Hamilton CM, Strader LC, Pratt JG, et al. The PhenX Toolkit: get the most from your measures. *Am J Epidemiol*. In press. (doi:10.1093/aje/kwr193).