Published in final edited form as:
JAMA. 2010 November 24; 304(20): 2290–2291. doi:10.1001/jama.2010.1686.

Enhancing the Feasibility of Large Cohort Studies

Teri A. Manolio, M.D., Ph.D.[1] and **Rory Collins, F.Med.Sci.**[2]

[1]Office of Population Genomics, National Human Genome Research Institute, Bethesda, MD

[2]Clinical Trial Service Unit and Epidemiological Studies Unit, University of Oxford, Oxford, UK

The identification of many hundreds of genetic variants associated with complex diseases and their potential for interactions with environmental factors have increased the need for prospective cohort studies involving several hundred thousand participants. [1,2]. Costs of such studies under conventional funding models are high because typically they are conducted by consortia of academic centers each responsible for recruitment, examination, and follow-up of a subcohort of participants in its geographic area [3,4]. The costs and inefficiencies in 100-fold expansion of these standard models can be prohibitive; large studies should not be viewed simply as small studies made large. Rather, they require fundamentally different approaches in which minimizing cost is a primary consideration, and "process" proficiency to maximize efficiency is as important as scientific expertise.

UK Biobank is a large prospective study that relies on a centralized strategy for nearly all aspects of its conduct. (5) This strategy, which UK Biobank adopted after rejecting a decentralized approach due to excessive cost, has achieved exceptional efficiencies while retaining scientific rigor. The main phase of recruitment started in April 2007 after a successful integrated pilot of the recruitment and assessment processes. The recruitment target of 500,000 individuals aged 40-69 years was achieved in July 2010, about 18 months ahead of schedule and within budget. (5) The UK Medical Research Council and Wellcome Trust charity are the chief funders (5), and the total cost of recruitment and baseline assessments, along with establishing the sample and data storage infrastructure, was about $100M. Annual costs for the subsequent phase of health outcome follow-up and adjudication, as well as maintaining the sample store and developing the IT systems to facilitate use by researchers, are estimated to be about $7M. The success of the UK Biobank model may provide valuable lessons for the efficient operation of large prospective studies in the US.

To Centralize or Not?

Centralized models of recruitment, data collection, sample processing, and follow-up can increase efficiency while promoting standardization, thus meeting both process and scientific imperatives. Rather than establishing and maintaining dozens or even hundreds of individual assessment centers, each of which must become expert in all study aspects and conduct them in (ideally) identical fashion throughout the study, participant recruitment can be focused in specific locales for a set period and then shifted to other areas. Cost efficiency can be achieved if assessment centers are inexpensive to establish and dismantle, temporary staff can be rapidly hired and trained, sizeable numbers of interested participants can easily reach the site, and economical space with good transport links is available for the period needed to "exhaust" an area of willing participants.

Correspondence: Teri Manolio, M.D., Ph.D., Director, Office of Population Genomics, Senior Advisor to the Director, NHGRI, for Population Genomics, National Human Genome Research Institute, Building 31, Rm. 4B-09, 31 Center Drive, MSC 2154, Bethesda, MD 20892-2154, manolio@nih.gov.

In UK Biobank, it was possible to meet all these conditions while maintaining good quality data collection. About 6 assessment centers were in operation at any one time, each recruiting about 100 participants per day for about 6 months, and a total of about 20 centers were required to recruit the complete cohort. (5) By comparison with distributed models typical of smaller studies, daily transmission of the data to the UK Biobank coordinating center facilitated real-time monitoring based on stable estimates of expected means, variances, missing rates, etc., allowing important aberrations to be detected and corrected rapidly. Centralizing responsibility for such day-to-day operations in skilled project managers has thus freed UK Biobank investigators to focus on science rather than miring them in logistical details.

Centralized models may also have drawbacks. Experienced investigators may have well-functioning recruitment and follow-up systems within their communities, as well as an understanding of unique local conditions. Academic centers accustomed to operational leadership in their geographic area may feel disenfranchised, risking the loss of their scientific and logistical input. Encouraging such investigators to provide procedural insights and advice for a centralized approach may make optimal use of their expertise, while streamlining lines of operational responsibility. Collaborating investigators can also assume primary responsibility for specific centralized aspects of study operations, such as enhancing recruitment of under-represented groups, responding to participant or community concerns, or developing systems for assessment of health outcomes.

Challenges in Event Ascertainment

Identifying disease events after the baseline assessment is as critical to the success of prospective studies as recruiting the cohort in the first place. Ensuring effective follow-up has been a major reason for establishing semi-permanent, localized study centers closely associated with area hospital record systems. In settings with one or only a few major sources of care and comprehensive medical records, follow-up may be possible with minimal participant input through record surveillance alone. Regional or national health care systems, such as the British National Health Service (NHS), can greatly simplify follow-up for clinically-detected events, though this will miss asymptomatic outcomes and may misclassify those with incorrect clinical diagnoses. UK Biobank will be relying on NHS records for disease ascertainment with subsequent centralized adjudication, but US systems are fragmentary, non-standardized, and challenging to access. If centralization is to work optimally, remote ascertainment of disease outcomes without ongoing contact with participants will almost certainly be necessary. This presents a considerable obstacle to large cohort studies in areas without central health data systems, though email or internet re-contact may become increasingly feasible. Studies in the US will be greatly facilitated by the development of standardized electronic medical records nationwide [6]. Embedding participant recruitment in an infrastructure for follow-up, as in studies conducted through established health care systems in the US [7,8] necessarily constrains the population from which a study can draw and may limit the diversity of the resulting cohort (but not its generalizability).

High Response Rates or High Recruitment Rates?

A key consideration in limiting costs of large prospective studies is the vigor with which a high participation rate is pursued. Nationally representative surveys such as the NHANES [] and disease-specific studies such as the Cardiovascular Health Study [3] provide valuable population-based estimates of disease prevalence and incidence which require high response rates. Prospective cohorts need not, however, be representative of a population to be generalizable; for example, the British Doctors' study provided valuable insights on the

disease risks due to smoking for the general population [9] and the Framingham study has provided information about blood pressure and cholesterol that go well beyond that one small Massachusetts town [10]. If a cohort study focusing on disease risk associations has a sufficiently large base population and captures a diversity of exposures and backgrounds, the results can still be applicable to populations with different distributions of these exposures. For these reasons, UK Biobank chose to emphasize diversity rather than participation rates, and accepted yields of roughly 10%. (5) If enough potential invitees are available, substantial savings can be realized by not attempting to persuade undecided persons to join.

Conclusions

The need for large prospective cohorts to assess genetic and environmental factors reliably requires nearly unprecedented levels of cost-efficiency. The novel approaches successfully used by UK Biobank may not be directly transferable to all settings in the US, where infrastructures for recruitment and follow-up differ, and diversity and distances are greater. Careful assessment and piloting will be needed to assess the feasibility of such models in the US. Several large-scale US efforts are under way including major initiatives by Kaiser and the Department of Veterans Affairs. However, if these large US cohort efforts are to be successful, lessons learned from approaches used by the UK Biobank may help point the way.

Acknowledgments

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

References

1. Collins FS. The case for a US prospective cohort study of genes and environment. Nature. 2004 May 27; 429(6990):475–7. [PubMed: 15164074]
2. Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. Nat Rev Genet. 2006 Oct; 7(10):812–20. [PubMed: 16983377]
3. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, et al. The Cardiovascular Health Study: design and rationale. Ann Epidemiol. 1991 Feb; 1(3):263–76. [PubMed: 1669507]
4. The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. Control Clin Trials. 1998 Feb; 19(1):61–109. [PubMed: 9492970]
5. (www.ukbiobank.ac.uk)
6. Office of the National Coordinator for Health Information Technology, Health IT Strategic Framework. [3/29/10]. http://healthit.hhs.gov/portal/server.pt/gateway/PTARGS_0_11673_911160_0_0_18/HIT_Strategic_Framework_032410_update.pdf
7. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): Design, methods and recruitment for a large population-based biobank. Personalized Med. 2005; 2:49–79.
8. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008 Sep; 84(3):362–9. [PubMed: 18500243]
9. Doll R, Peto R, Boreham J, Sutherland I. Mortality from cancer in relation to smoking: 50 years observations on British doctors. Br J Cancer. 2005 Feb 14; 92(3):426–9. [PubMed: 15668706]
10. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J 3rd. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. Ann Intern Med. 1961 Jul.55:33–50. [PubMed: 13751193]