COMMENTARY

Merging and emerging cohorts

How best to study the effects of genes and environment on US health? In the first of two commentaries, **Walter C. Willett** and his co-authors argue that investing in existing studies is the most efficient approach. In the second, **Francis S. Collins** and **Teri A. Manolio** explain their support for a new national cohort.

Not worth the wait

In 2006, the United Kingdom initiated a national long-term health study of 500,000 middle-aged adults that will involve collecting DNA and other biological specimens[1]. Further cohorts are being considered elsewhere in Europe and Asia. Francis Collins (ref. 2 and page 259) has proposed a similar national cohort of several hundred thousand North Americans to enable future studies of the genetic basis of human diseases and individual susceptibility to environmental factors. The cost is estimated to be US$3 billion or more[3].

We are concerned that results from a new cohort would not be available for at least ten years, as five years would be needed for funding, planning and enrolment, and another five for following up even the earliest analyses of the most common diseases; results for most cancers would take longer. We believe that a strategy using existing cohort studies should also be considered. We argue here that this approach can achieve the anticipated objectives of the national cohort more rapidly and more cheaply, and with similar scientific validity.

Already available

Much of what is known about the causes of cancer, heart disease and other illnesses has arisen from epidemiological studies, especially cohort studies. Prospective cohorts, such as the Framingham Heart Study, have had success in identifying key determinants of major diseases, including smoking, physical activity, occupational exposures and adverse effects of medications[4]. In the late 1970s, several much larger cohort studies started to assess lifestyle factors, with many also collecting and storing blood samples for analyses of DNA and plasma[5].

The Table overleaf lists large US cohorts, which total nearly 1,400,000 participants with biologic specimens available for more than 800,000 of them. There are also several ongoing medium-sized cohorts, with typically 5,000–20,000 available biospecimens, which focus on cardiovascular disease[6]. These cohorts also record detailed biological measurements and repeated measures of atherosclerosis.

GETTY

Although most large or medium cohort studies were initially designed to investigate cancer or cardiovascular disease, each routinely ascertains information on other diseases and lifestyle. In the Nurses' Health Study, for example, published endpoints include stroke, diabetes, cataracts, asthma, gallstones, multiple sclerosis, Parkinson's disease, phobic anxiety and quality of life, among many others. The ability to study multiple endpoints within the same population, at modest additional cost, is invaluable in formulating public-health and clinical guidelines, because some environmental exposures may decrease the risk of one disease while increasing the risk of another.

Each existing cohort has tracking systems enabling retrieval of the biospecimens for assay as emergent hypotheses arise. Most cohorts collect comprehensive information on the participants at enrolment, including data on demographic, medical, familial, occupational and lifestyle (such as alcohol intake and diet) characteristics. Such information is usually limited to adult experiences, but lifetime histories are solicited for some exposures such as smoking.

Moreover, participants have already provided informed consent for the use of their biographical data and biospecimens for research. Institutional review boards have approved each study and monitor new ethical issues as they arise. Many investigators already have data-sharing policies in place for collaborations. Reconsent is sometimes needed with existing or new cohorts to accommodate new ethical issues, but this is more easily obtained when a trusting relationship already exists between participants and investigators.

Pooling power

Individual cohorts have already provided key clues to disease aetiology and prevention, but combining data from multiple cohorts offers even greater potential when studying the interplay between environmental and genetic factors. In the past, when findings from individual cohorts seemed inconsistent or inconclusive, data from several cohorts have been combined, sometimes in meta-analyses that attempt to overcome problems of small sample size. But meta-analyses of published reports can also be unsatisfactory because of publication bias towards positive results and because of results that are hard to combine statistically owing to differences in analysis and reporting of data.

Instead, groups of investigators are now collaborating to analyse the primary data from their studies jointly. For example, the Pooling Project of Cohort Studies of Diet and Cancer, which includes more than 25 cohorts from the United States, Europe and Asia with more than 2 million men and women (of which half have biomarkers), has published a series of pooled data on diet and cancer[6,7]. When these data are analysed in a standard manner the findings are highly consistent.

A similar consortium, coordinated by the National Cancer Institute (NCI), is systematically genotyping samples and combining results for breast and prostate cancer to investigate gene–environment and gene–gene

interactions[8,9]. So far, this consortium includes about 10,000 cases of both cancers, along with matched controls from the same cohorts, which allows for robust exploration of candidate genetic pathways. To attain these same numbers with a US national cohort would require following up 500,000 women for 20 years and 500,000 men for 10 years — or 15 million person-years of follow-up.

As an example of the NCI consortium's ability to address new leads rapidly, within weeks of the publication of a polymorphism for prostate cancer risk[10], the consortium was able to confirm this finding in seven individual cohorts. Indeed, whole-genome scanning of 10,000 cancer cases and 10,000 controls could identify almost all common genetic variants even weakly associated with breast and prostate cancer at an estimated cost of $5–10 million.

Existing large US cohort studies with biospecimen repositories containing blood and/or DNA samples

Study	Year biospecimen collection began	Total cohort size	No. with stored biological samples
Health Professionals Follow-Up Study	1986	52,000	30,000
Nurses' Health Study I	1989	122,000	63,000
Washington County Study	1989	33,000	33,000
Women's Health Study	1992	40,000	28,000
Women's Health Initiative	1993	162,000	162,000
NCI PLCO Study	1994	155,000	70,000
Nurses' Health Study II	1996	116,000	60,000
American Cancer Society CPS-II LifeLink Study	1998	184,000	109,000
Multiethnic Cohort Study	1996	215,000	80,000*
Vitamins and Lifestyle (VITAL) Cohort	1999	78,000	54,000
Agricultural Health Study	1999	90,000	35,000
Southern Community Cohort Study	2002	90,000*	80,000*
Black Women's Cohort Study	2005	59,000	41,000*
Total	**-**	**1,396,000***	**845,000**

*Expected totals upon completion

Dispelling myths

Certain limitations of the existing cohorts have been raised[2,11]. One oft-repeated concern is that existing cohorts are not representative of the US population. However, representativeness is not needed to determine the relation between genetic or environmental factors and disease risk; for example, the effects of smoking were first revealed in a study of British doctors. In many cases, cohort studies are purposefully nonrepresentative to maximize the quality of data or to emphasize the contrast between environmental exposures. A cohort of Seventh-Day Adventists, for example, allows the investigation of high consumption of soya products, and studies of nurses or doctors provides insight into long-term behaviours thought to be healthy but under-represented in the general population.

A related myth is that population subgroups, whether racial, gender, or religious, must be represented in proportion to their prevalence in the general population in any single study to be confident that the results apply to these subgroups. But what is really needed is a subgroup large enough to examine the exposure and disease association within that subgroup. The cohorts with biospecimen repositories listed in the Table include large numbers of African-Americans, Hispanics and other minority groups. In our view, the US National Institutes of Health (NIH) research portfolio already has adequate statistical power to investigate genetic variations and environmental factors within most major ethnic subgroups except for Asians.

Another criticism of existing cohorts is that self-reported measures of diet, physical activity, and other environmental exposures are less useful than more time-intensive and costly objective measurements. But because diet and physical activity vary substantially over time for most individuals, even an objective measure of short-term exposure (which rarely exists) can be inferior to a questionnaire that solicits less precise information over a longer period. Also, biomarkers using plasma, red blood cells, DNA and nail samples are now being used to assess many dietary and other environmental factors, and can be integrated into existing cohorts. For many biomarkers, the existing specimen repositories will be invaluable because millions of person-years of follow-up are already available.

Layered approach

Because most existing cohorts enrolled middle-aged or older adults, information is sparse on childhood and early adult years. Young age may be a critical window of exposure to cancer risks, for example[12]. Ultimately, cohorts established during childhood will be more informative, but decades of follow-up will be required for information on adult diseases. Although not perfect, some existing cohorts seek information on early-life exposures based on interviews with the adult participants.

In addition, the Nurses' Health Study II has enrolled around 25,000 offspring of existing participants between the ages of 10 and 14; some have already been followed up for ten years. Biological specimens are not yet available, in part because of the cost and complexities regarding informed consent from children, but DNA could be collected at age 21. This cohort is probably large enough to study most common outcomes, say type 2 diabetes or asthma, but not less-common outcomes such as specific cancers. We believe a layered approach makes sense because most common outcomes, such as hypertension, can be addressed by detailed studies of medium-size cohorts, and less common outcomes, such as cancer, can be studied more cheaply with less-intensive approaches using large cohorts.

Substantial investments have already been made, mainly by the NIH, to create cohorts with more participants than the proposed new cohort, and sufficient follow-up to conduct powerful genetic analyses for major cancers and other health endpoints. Although the existing cohorts have some gaps in data and specimens, these can be filled with relatively modest investments. In an era of limited research funding, a wise strategy would be first to use existing resources more efficiently before embarking on an extraordinarily expensive new cohort, which will provide little or no return for the next decade or more. ■

Walter C. Willett is in the Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts 02115, USA; his co-authors are William J. Blot, Graham A. Colditz, Aaron R. Folsom, Brian E. Henderson and Meir J. Stampfer. For additional information, e-mail walter.willett@channing.harvard.edu.

1. Watts, G. *BMJ* **332,** 1052 (2006).
2. Collins, F. S. *Nature* **429,** 475–477 (2004).
3. Spivey, A. *Environ. Health Perspect.* **114,** A466–A467 (2006).
4. Samet, J. & Munoz, A. *Epidemiol. Rev.* **20,** 1–14 (1998).
5. Langholz, B., Rothman, N., Wacholder, S. & Thomas, D. C. *Monogr. Natl Cancer Inst.* **26,** 39–42 (1999).
6. National Heart, Lung and Blood Institute population studies database. http://apps.nhlbi.nih.gov/popstudies/ (accessed 24 August 2006).
7. Smith-Warner, S. A. *et al. Am. J. Epidemiol.* **163,** 1053–1064 (2006).
8. Hunter, D. J. *et al. Nature Rev. Cancer* **5,** 977–985 (2005).
9. Feigelson, H. S. *et al. Cancer Res.* **66,** 2468–2475 (2006).
10. Amundadottir, L. T. *et al. Nature Genet.* **38,** 652–658 (published online 7 May 2006).
11. Secretary's Advisory Committee on Genetics, Health and Society. http://www4.od.nih.gov/oba/SACGHS/public_comments.htm (May 2006).
12. Land, C. E. *et al. Radiat. Res.* **160,** 701–717 (2003).

Acknowledgment The authors acknowledge the helpful suggestions of D. Hunter.