Necessary but not sufficient

The proposal advocated in the preceding Commentary by Willett *et al.*[1], namely to extend existing cohort studies rather than start a new large-scale prospective study from scratch, has many merits. Indeed, a US National Institutes of Health (NIH) study group that assessed the pros and cons of various models in 2004 considered this option in some depth, and their report[2] made many of the same points.

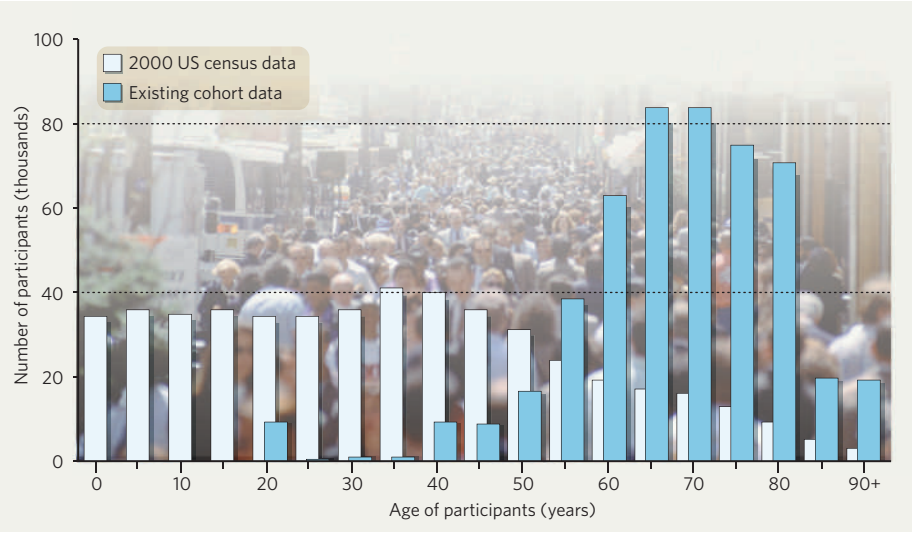
Certainly, assembling existing cohorts into a large consortium would provide a powerful resource for investigating genetic and environmental factors in health and disease. The argument that this method is likely to be less costly than a new cohort, and would yield results more quickly, carry considerable weight. But Willett *et al.* do not address all of the suboptimal aspects of this approach. Those should be clearly noted, lest expectations of such a consortium exceed what it is likely to deliver.

First, there is the issue of standardization. Phenotypic measures used by the existing cohorts, although standardized within cohorts, have not followed uniform procedures across studies, and so there will be significant challenges to merging data from different studies in a valid way. Moreover, key environmental exposures or risk factors will almost certainly differ systematically across cohorts. Combining studies that were focused on specific population subgroups will therefore introduce biases that can be corrected only by limiting the analysis to the lowest common denominator of valid, unbiased exposures.

Second, the reliance on legacy studies fails to take advantage of new tools for measuring dietary intake, physical activity and environmental exposures (as are now being supported through the NIH Genes and Environment Initiative[3]), because many of these measurements — such as precise ambulatory data — cannot be made on stored biospecimens.

Third, representation has been a major concern driving the national cohort proposal. Despite recent attempts to improve representation of minorities and socioeconomically disadvantaged participants in newer cohorts, the proportions are still far below their representation in the US population. There is also substantial under-representation among men and participants from the south, as well as those with lower levels of education, although these might be addressed somewhat by statistical adjustments.

Fourth, under-representation of people younger than the age of 50 is substantial in these existing cohorts (see Figure) and will only get worse with time. If we wish to address complex disease risk across lifespan, we need to study diseases developing in adolescence and young adulthood, such as asthma, autoimmune disease and major psychoses. Even the investigation of mid-life diseases would be limited by lack of stored biospecimens in earlier life.



GETTY

Comparison of an estimated distribution of a 500,000-person cohort based on existing cohort data[2], with the 2000 US census.

Finally, the full value of a large-scale cohort study will depend on free and open access to the data by all qualified investigators. This may be difficult to achieve with a combination of existing cohorts, given the expectations of current investigators about control of the data, and consent limitations by existing study participants.

More ways than one

There is no question that a new cohort study would require many years to implement and to generate results, although useful findings would be available on more common diseases within five years of cohort recruitment[4]. We agree with Willett *et al.*, therefore, that it is reasonable in the interim to seek ways to form consortia of existing studies. But these two models need not be thought of as mutually exclusive.

We must also recognize that environmental exposures (including emerging infections) and preventive or therapeutic interventions will probably change dramatically in the next two decades. Limiting our research enterprise to the exclusive study of existing cohorts, especially without collection of new risk information and recruitment of participants under-represented in existing studies, may ultimately jeopardize our ability to address these evolving health risks in an epidemiologically rigorous manner.

Admittedly, this discussion remains hypothetical, because serious budgetary challenges make a new national cohort an unlikely prospect at the present time. Some may wonder whether the United States can afford both an expansion of existing cohorts and a new national cohort. We believe the real question is whether it can afford not to do both, given the enormous and growing healthcare costs of complex diseases. Finding the genetic causes of even one of these diseases could potentially save billions of dollars in medical costs if appropriate preventive interventions can be developed. Despite the fiscal realities, therefore, we must continue to make the case both for a merging of cohorts now, and the founding of a more rigorously designed national cohort in the future when funds are available. Although recognizing the massive uncertainties in budget situations and priorities, we believe that future generations will wonder why we didn't try as hard as possible to get both of these kinds of studies underway. ■

Francis S. Collins and Teri A. Manolio are at the National Human Genome Research Institute, National Institutes of Health, 31 Center Drive, Bethesda, Maryland 20892-2152, USA.

1. Willett, W. C. *et al. Nature* **445,** 257–258 (2007).
2. www.genome.gov/Pages/About/OD/ReportsPublications/PotentialUSCohort.pdf
3. NIH Genes and Environment Initiative www.gei.nih.gov/
4. Manolio, T. A., Bailey-Wilson, J. E. & Collins, F. S. *Nature Rev. Genet.* **7,** 812–820 (2006).

Acknowledgements We acknowledge A. Guttmacher, E. Harris and L. Rodriguez for their advice.