Building a Consortium of Cohorts – Cohort Identification and Participant Recruitment

Rebecca Baker Ph.D., NIH; Eric Boerwinkle, Ph.D., University of Texas Health Science Center, (Co-chair); Greg Burke M.D., M.Sc., Wake Forest School of Medicine; Rory Collins, F.MedSci., University of Oxford; Michael Gaziano M.D., M.P.H., VA Boston Healthcare System, Michael Lauer, M.D., NHLBI (Co-chair); and Teri Manolio M.D., Ph.D., NHGRI.

Publication note: This document is an invited "white paper" created as background information for participants at a workshop held on February 11-12, 2015. It represents the opinions of the authors, and is intended to help inform the planning process for the Precision Medicine Initiative research cohort.

Draft Report

Executive Summary

The NIH proposes to create a national cohort of at least one million Americans – committed to participate in research – to advance our understanding of heath and disease.[1] The national cohort will be unprecedented in scope, and will recruit expertise from multiple sectors to make genomic, environmental, lifestyle, and electronic medical record information available to investigators. To ensure that this initiative will meet its enormous promise in a timely manner, it would be best to build upon and collaborate with a robust platform of existing cohorts.

By assembling existing cohorts into a large consortium of cohorts, with a central infrastructure, NIH could harmonize data types; enhance data collection; achieve economies of scale; and provide a resource for addressing new scientific questions, developing new technologies, gathering and sharing heath information, and enabling new patient-powered models that will drive biomedical research and healthcare. Together, these new opportunities are likely to advance precision medicine and to improve the health of all Americans.

Introduction

A working group of investigators, from both NIH and the extramural community, was assembled to identify problems related to assembling and enhancing a consortium of cohorts, and to outline concrete steps to meet these challenges. The working group members brought expertise in cohort design, management, and recruitment, as well as in research methodologies related to classical epidemiology and modern-day "mega-epidemiology" that leverages "big data" and rapidly evolving high-throughput technologies.

Background

Precision medicine aims to tailor therapies to an individual's unique traits – be they genetic, molecular, environmental, psychosocial, or economic. Technologic advances including a reduction in the cost of DNA sequencing, especially complete genome sequencing, combined with an expansion of phenotype information to include digital data sources, mobile health, and electronic health records, offer new opportunities for the NIH to advance precision medicine. With support from the President[1], NIH is planning a Precision Medicine Initiative to assemble a large national research cohort of at least one million Americans who will agree to share their whole genome sequence and other information. The cohort will integrate genomic, clinical and other health-related information into a framework accessible to and useful for researchers investigating a broad range of diseases, including cancer, and medical questions, including gene-environment interactioninteractions. Information from the cohort, combined with patient-powered research approaches and cutting-edge technologies, will help develop new disease prevention strategies, novel therapeutics and medical devices, and improve the effectiveness of treatments by tailoring them to individual characteristics.

Because only a small proportion of the cohort will develop any particular condition, the new cohort initiative will require a large number of participants. A study of at least one million is needed to provide researchers with the broad range of health outcomes, genotypes, and exposures necessary to discern the subtle but important individual contributions that will inform precision medicine. The most informative existing data sources will come from established longitudinal cohort studies, where participants have already been recruited, enrolled and measured for multiple risk factors, and followed over time for ascertainment of outcomes of interest. However, rather than a mere collection of existing cohort studies, the national cohort will provide enhanced detailedin-depth phenotyping, and create enhanced opportunities for study participation by both the subjects and scientific community. In this way, the cohort aims to address a wide range of biomedical questions not otherwise possible in the individual cohorts or via meta-analysis. Additional data on participant exposures, risk factors, enhanced phenotypes, and well-described patient-centered health outcomes will come from continued follow-up, as well as through electronic health records, mobile digital sensors, web-based communication portals, and disease registries.

The optimal cohort study will be large in scale; will include data on a wide range of diseases and health outcomes, including patient-reported outcomes; will be designed to serve the interests of both participants and researchers; and will collect, standardize, link, and share data efficiently.[2,3,4] The cost of recruiting such a sample from scratch is prohibitive; furthermore, because it takes time to recruit subjects and to wait for outcome events to accrue, there will be an unacceptably long delay between inception and meaningful research output. We therefore see as a cost-effective solution combining existing longitudinal studies. In the short-term, the pooled cohorts offer scientists a resource for interesting analyses; more importantly, for the long-term the pooled cohorts offer a "head start" for prospective investigations. Another related approach would be to leverage cohorts that are increasingly being recruited through regional health-care providers and professional societies, which often have access to large numbers of people with a broad range of health and electronic health record data.[3,5] Participants in existing cohorts and professional society registries could be asked to join new participants in the new, enhanced cohort.

A preliminary NIH inventory report identified 50 large-scale research cohorts, including 42 federally funded studies, 6 hospital-based cohort studies, and 9 hospitals and healthcare systems with research oriented databases of their patients. Together these comprised approximately **12.3 million individuals** enrolled across the 65 studies, including **6.8 million** people taking part in the studies receiving NIH funding. Genotyping and broad data sharing (e.g. through the NIH database of Genotypes and Phenotypes dbGAP) of at least some samples has occurred in 42 studies. Although these cohorts have each been leveraged for research studies, their data and samples have not been combined and harmonized into a large-scale accessible research-quality database and resource. The assembly of a consortium of cohorts from these individual collections is clearly possible. Additional detailed characterization of extant cohorts should be undertaken to help in the planning.

The working group believes that it would be ideal if assembly of and recruitment into a national cohort were complete in 5 years. Participants will have the option to contribute their health information on a lifelong or periodically renewable basis. Short-term goals of the cohort study might include analyses that can be performed with baseline data and follow-up to date, along with information provided by participants through surveys and mobile devices. In the longer term, enhanced phenotype measures and electronic health records collected on newly recruited or consented participants will empower a unified cohort with the durability necessary to perform new types of analyses and advance precision medicine.

Barriers and challenges to building a large national cohort:

The assembly of a new national cohort of such unprecedented size is a bold undertaking, and will face significant challenges. The numerous NIH-supported research cohorts benefit from the enthusiasm of local institutions, local participants and local investigators, and have offered valuable answers to a wide array of scientific questions. These studies are richly phenotyped, contain valuable longitudinal data, incorporate broad ethnic diversity, and have invested in the necessary materials, including DNA samples and IRB approvals, to move forward with DNA sequencing and broad data sharing. However, a common infrastructure to bring together information from these various studies is lacking. Each individual cohort study has pursued different approaches to data gathering, sharing, and the involvement of research participants, including provisions for contact and consent. Therefore, efforts within these studies may be inefficiently duplicative and often present barriers to collaboration. The creation by NIH of a national cohort of at least one million Americans would face challenges such as:

- **Expense** – conceptualizing and starting a project, especially on this scale, is expensive; policy makers and the public will expect "early returns" on such a large investment. Furthermore, the cost savings realized through leveraging and pooling existing cohorts may be overestimated.
- **Time** – Longitudinal studies of chronic disease outcomes span many years (sometimes decades) to allow for a robust number of endpoints to occur. It is not unusual for an epidemiological cohort to make substantial discoveries until years after initial enrollment.
- **Feasibility** – The Federal Government's ability to efficiently launch a large-scale cohort study will be called into question.
- **Contact** – Existing cohorts are heterogeneous with respect to permission for data sharing and the need for researchers to re-contact/consent participants to join the study.
- **Demographics** – Existing US research cohorts do not completely represent the American population, and do not mirror projected demographic changes in the American population.

        - Age: While older people may be over-represented in existing cohorts[6], they are also more likely to develop disease; enrolling large numbers of younger people may be inefficient
        - Race/Ethnicity: There is a need for additional representation for multiple non-European ancestry populations
    - **Privacy** – Participants may have suspicions about the Federal Government "prying" into private health data, and concerns about the security of their individual data and health records.
    - **Dynamic Technologies** – Administrative-claims, digital and smart-phone technologies to track participants over time and space are rapidly evolving so that chosen approaches may rapidly become obsolete. In addition, there is growing tension between these technologies and the desire for privacy.
    - **Scope** – Sufficient sample size required to capture small proportion of people with a specific disease or genotype.
    - **Competition** – Other countries (e.g. the United Kingdom[7], Germany and Denmark[8]) have already launched large national cohort studies, and any new effort will need to be fully justified.
    - **Perceived need** – The added benefit of an expensive new cohort may not be immediately clear, given that many large cohorts already exist (and others are underway – a non-comprehensive sampling includes PCORnet and the Study of Latinos). The NIH and investigators will need to identify and present "use cases" that justify the added effort and expense. Critics might ask, "What do you expect to learn in the next year, the next 5 years, and the next 15 years that could not have been learned from ongoing efforts and investments?"
    - **Coordination, transparency, and governance** – Necessary information is not readily available and useful to investigators:
        - Fragmented electronic health records and claims data,
        - Fragmented data platforms,
        - Fragmented health care system in which tracking people across space and time is inherently problematic, e.g. state by state carrier boundaries.
    - **Incentives** – Investigators have invested considerable financial, emotional, and intellectual capital in in existing cohorts, and may not be willing to fully share the data these cohorts contain. Researchers and research participants will need to see the newly integrated and greatly expanded cohort as a resource of added value, not a parochial threat, to their own personal and professional interests.

Approaches to overcome barriers and challenges:

To best overcome these challenges, the NIH Precision Medicine cohort should take the form of a consortium of cohorts, with a centrally coordinated infrastructure overseen and perhaps directly managed by NIH. By taking full advantage of existing cohorts, the NIH research cohort could provide a coordinated interdisciplinary approach to address scientific questions, achieve economies of scale, create opportunities for collaboration, and accelerate the pace of research and the implementation of new methods.

- **Expense, time, and feasibility** – Construction of a large national research study, as a synthetic cohort assembled from numerous existing research cohorts, would leverage existing resources; many cohorts already have generated extensive genomic data.
  - Existing cohorts can be combined with a focused recruitment of new participants.
  - Existing cohorts and biobanks, especially those in which "recruitment" is minimal because consent has already been obtained to share data for further research (e.g. The Atherosclerosis Risk in Communities and Million Veterans Project)
  - Existing electronic health records (e.g. Partnership of Mayo with Optum Laboratories)
  - Existing data registries (e.g. National Cardiovascular Data Registries)
  - Existing digital communities (e.g. PatientsLikeMe, Health eHeart).
- **Technology** – The development of ongoing internet and cloud-based services rather than individual apps and unidirectional data sharing may enable better long-term gathering of new data types from mobile, home-monitoring, and electronic health records to supplement doctors' examinations.
  - **Scope –** to maximize the scientific value of a cohort of at least one million, the cohort could be preferentially enriched for designs that reduce sample size requirements -- e.g. Family studies, isolated communities (e.g. Amish), and phenotypes "close" to the gene level, such as metabolomics.
- **Coordination** – A central coordinating center could foster harmonization of existing data, as well as facilitate the testing and evaluation of new data collection methods across multiple cohorts, both established and new.
  - A steering committee, composed of investigators, could advise NIH on policy, management and scientific direction of the cohort consortium provide infrastructure support to enhance use of limited resources, and reduce the marginal cost of new studies.
- **Enhancing Incentives –** Access to full-genome sequencing and other cutting edge technologies, such as internet-based questionnaires (cognition, mood, pain, quality of life), novel devices and apps, and advanced imaging may provide incentive for investigators leading existing cohorts to join the national cohort.

Opportunities for building a large national cohort

The creation of a U.S. Research Cohort will be aided both by the solid foundation laid by NIH-funded researchers and existing research studies, as well as by new technological developments in the types of data gathered about research participants. Deep phenotypes associated with existing cohorts could be complemented by new categories of information including mobile technology and electronic health records. Similarly, longstanding NIH efforts to promote data sharing and collaboration could be used as a framework for the development of new technology and tools. By leveraging ongoing technological, social, economic and policy trends, NIH can encourage the optimal utilization of NIH resources in a large cohort study. Opportunities for the cohort study include:

- Adoption and expansion of smart-phone technologies will enable the collection of new data types from mobile, home-monitoring, and electronic health records to supplement doctors' examinations.[9]
- Increasing integration of health care
- Increasing penetration of electronic health records[10]
- Efforts of NIH Institute and Centers -- e.g. NHLBI/NHGRI Whole-Genome Sequencing Project
- A new research culture in which data sharing is increasingly accepted and recognized as productive.[11]
- NIH Big Data to Knowledge (BD2K) initiative, under direction of Associate Director for Data Science Phil Bourne.
- Lessons from examples of successful consortia of cohorts -- e.g. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium.

The Enhanced U.S. Precision Medicine cohort will enable researchers to: (in expeditious and efficient manners):

- Study uncommon diseases;
- Identify genetic variants protective against disease;
- Explore disease penetrance and phenotypic implications for variants considered essential for life;
- Study new diseases and health outcomes (mental health, quality of life and well-being);
- Establish a platform for intelligent clinical trials that can be conducted at relatively low marginal cost;
- Employ and test new technologies for tracking health information;
- Empower a community of research participants to both contribute to research studies and understand their own health and disease.

Summary and future goals

The investments in U.S. health research made by the NIH have already provided a framework for creating a major novel research resource of at least one million Americans, containing genomic, clinical and other health information. An enhanced consortium of cohorts could build upon this framework for future NIH research and education initiatives, such as the creation of a scientific commons whereby a broad array of investigators can access the data they need (and are entitled to) to answer their specific scientific questions. The U.S. Research Cohort could also provide a platform for researchers to conduct rapid and efficient randomized trials that use cohorts and registries as ready-made platforms.[12] By providing a resource for these cutting-edge advances, the national cohort, as a consortium of cohorts, will help the US remain the leader in biomedical research, accelerate the transition to personalized medicine, and improve the health of US citizens.

We look forward to a dynamic dialogue at NIH's February 11 workshop, after which we plan to update this White Paper with suggested concrete "next steps."

References:

1. Collins FS, Varmus H. A New Initiative on Precision Medicine. *N Engl J Med*. 2015;

2. Collins FS. The case for a US prospective cohort study of genes and environment. *Nature*. 2004;429:475–477.

3. Manolio TA, Weis BK, Cowie CC, Hoover RN, Hudson K, Kramer BS, Berg C, Collins R, Ewart W, Gaziano JM, Hirschfeld S, Marcus PM, Masys D, McCarty CA, McLaughlin J, Patel A V, Peakman T, Pedersen NL, Schaefer C, Scott JA, Sprosen T, Walport M, Collins FS. New models for large prospective studies: is there a better way? *Am J Epidemiol*. 2012;175:859–866.

4. Lauer MS. Time for a creative transformation of epidemiology in the United States. *JAMA*. 2012;308:1804–1805.

5. Khoury MJ, Lam TK, Ioannidis JP, Hartge P, Spitz MR, Buring JE, Chanock SJ, Croyle RT, Goddard KA, Ginsburg GS, Herceg Z, Hiatt RA, Hoover RN, Hunter DJ, Kramer BS, Lauer MS, Meyerhardt JA, Olopade OI, Palmer JR, Sellers TA, Seminara D, Ransohoff DF, Rebbeck TR, Tourassi G, Winn DM, Zauber A, Schully SD. Transforming epidemiology for 21st century medicine and public health. *Cancer Epidemiol Biomarkers Prev*. 2013;22:508–516.

6. Collins FS, Manolio TA. Merging and emerging cohorts: necessary but not sufficient. *Nature*. 2007;445:259.

7. Collins R. What makes UK Biobank special? Lancet. 2012;379:1173–1174.

8. Frank L. Epidemiology. When an entire country is a cohort. Science. 2000;287:2398–2399.

9. Topol E. The Patient Will See You Now: The Future of Medicine is in Your Hands. Basic Books; 2015.

10. Wright A, Henkin S, Feblowitz J, McCoy AB, Bates DW, Sittig DF. Early Results of the Meaningful Use Program for Electronic Health Records. *N Engl J Med*. 2013;368:779–780.

11. Paltoo DN, Rodriguez LL, Feolo M, Gillanders E, Ramos EM, Rutter JL, Sherry S, Wang VO, Bailey A, Baker R, Caulder M, Harris EL, Langlais K, Leeds H, Luetkemeier E, Paine T, Roomian T, Tryka K, Patterson A, Green ED. Data use under the NIH GWAS Data Sharing Policy and future directions. *Nat Genet*. 2014;46:934–938.

12. Lauer MS, D'Agostino Sr. RB. The randomized registry trial--the next disruptive technology in clinical research? *N Engl J Med*. 2013;369:1579–1581.