Design Considerations for a Potential United States Population-Based Cohort to Determine the Relationships among Genes, Environment, and Health: Recommendations of an Expert Panel

Table of Contents

Design Considerations for a Potential United States Population-Based Cohort to Determine the Relationships among Genes, Environment, and Health: Recommendations of an Expert Panel

I. BACKGROUND AND RATIONALE

The completion of the Human Genome Project provides an unprecedented opportunity to define the genetic and environmental contributions to health and disease. Identifying genetic and environmental factors that influence health, disease, and response to treatment is essential to developing approaches to reduce disease burden - the primary goal, of course, of biomedical research. Several recent developments suggest that progress in this area could be quite rapid. In particular, the sequencing of the human genome, increasing information about its function, and the exploration of human genetic variation through the International Haplotype Map project, are providing powerful research tools for identifying variants that contribute to common diseases. In addition, progress in measuring non-genetic factors and environmental exposures promises to extend the range of observational epidemiologic investigation. Recognition is growing that environmental change interacting with genetic predisposition has produced most of the recent epidemics of chronic disease, and may thus hold the key to reducing such health problems in the future. Together, these developments present exciting opportunities to address key unanswered questions related to the complex contributions of genes, environment, gene-gene, and gene-environment interactions to health, with potentially major consequences for prevention, diagnosis, and treatment.

Many avenues of research must be pursued to realize the full potential of the Human Genome Project, including molecular investigations into the structure and function of the genome, characterization and annotation of genome sequence variation, examination of gene function and regulation in experimental systems, elucidation of these effects in animal models, and research into the broad array of societal implications of genomics. Identification of genetic variants related to complex diseases (diseases influenced by many genetic and environmental factors working together) requires correlation of genotypic and phenotypic data in human populations. Such studies are most commonly carried out using the case-control method, in which genetic and environmental data are collected from persons with specific diseases or conditions and compared to those free of disease. Although case-control studies are of great value in suggesting potential etiologic factors, they cannot provide information on predictive biomarkers, are prone to important biases related to case ascertainment, and often involve incomplete or biased assessment of risk modifiers or gene-environment interactions.

Rigorous and unbiased conclusions about disease etiology and population impact require population-based cohort studies, in which representative samples of a population are followed prospectively for development of specified endpoints. Phenotypic and environmental information collected in a standardized and unbiased manner is crucial to such efforts. Large-scale cohort studies are under discussion or already underway in the United Kingdom, Iceland, Estonia, Germany, Canada, Taiwan, and Japan. While such projects are likely to be powerful engines for research, other nations' projects alone will not meet U.S. needs, due to inadequate representation

of U.S. minority groups that bear disproportionate burdens of disease, the limited potential for research access to data and biological materials, and substantial international differences in environment, lifestyle, and health care. Were a U.S. study to be considered, however, it would be best for U.S. investigators to work closely with international studies so that results can be compared across populations. This would permit examination of the key issues of generalizability of findings and modification of relationships by underlying genetic backgrounds and environmental exposures that are unique or disproportionately common in one setting or another.

For these reasons, the National Human Genome Research Institute, in collaboration with several other NIH Institutes, commissioned a group of experts in genetics, epidemiology, biostatistics, and ethical, legal, and social issues in genetic research to examine the scientific foundations and broad logistical outlines of a hypothetical U.S. cohort study of genes, environment, and health (Appendix 1). Although funding for such an endeavor has not been identified, and may never be identified, carefully outlining and considering the goals and key design aspects of such a study was deemed of high scientific importance. Exploration of these issues permits a rigorous assessment of what existing and potential future U.S. studies, as well as other nations' efforts, can contribute to understanding the complex contributions of genes, environment, and their interactions to health and disease in the United States. The recommendations of this expert panel are summarized below.

II. GOALS

II.A. Potential Objectives and Research Questions

Any large U.S. cohort study of genes, environment, and health should be designed to address an ambitious goal: to ascertain and quantify all of the major environmental and genetic causes of common illnesses, setting the stage for a future of better preventive medicine and more effective therapy. Such a study could examine the environmental exposures, genetic risk factors, lifestyle, and medical experiences of a cross-section of America of unprecedented size and scope. Using the most advanced research technology and rigorous methods to protect individual privacy, hundreds of thousands of Americans could be enrolled and engaged to partner with researchers to identify the causes of common diseases across a diversity of ages, geographic regions, and ethnicities.

The ambitious goals of such a study would fall into four categories, each of which could include significant deliverables:

1. Information that would be derivable very soon after enrollment of participants
    - Population prevalence of known genetic variants
    - Population prevalence of common diseases and environmental exposures
    - Associations of genetic variants and environmental factors with traits and conditions present at baseline
2. Information that would be derived from incident cases as the cohort progresses
    - Determination of quantitative, unbiased risk of major genetic and environmental

susceptibility factors for common, complex disease
- Identification of major gene-by-environment and gene-by-gene interactions
- Identification of biomarkers that represent early indicators of disease

3. Technology
   - Development of sophisticated methodology for collection of large-scale phenotypic and environmental data
   - Development of sophisticated methodology for data mining and statistical analysis appropriate for studies of genetic and environmental influences
4. Access
   - Wide availability of this resource to the scientific community

Goals related to population prevalence, cross-sectional associations, methodological development, and resource establishment could be accomplished early in such a project (within the first few years), while assessment of risk and identification of interactions would require sufficient follow-up for new cases to accrue. Persons with prevalent disease should be included at entry (see Section IV.C) to allow early cross-sectional analyses of genetic and environmental associations with diagnosed disease. To ensure standardized phenotypic definition and unbiased ascertainment, however, the most reliable analyses would be those of incident cases occurring after entry to such a study. Although thousands of cases of the most common diseases would be expected to occur within the first few years, other conditions would require longer follow-up for sufficient numbers of cases to accrue. Conditions to be ascertained should be based on the leading causes of disability adjusted life years (DALYs) lost, death, and hospitalization, as well as diseases with the highest incidence and prevalence (Table 1).

II. B. Relationship to Existing U.S. Cohort Studies

NIH has supported a large number of population-based U.S. cohorts, involving more than 1.5 million people, some of whom have been under study for many decades. Although most of these cohorts focus on a single condition or group of conditions, such as cancer or heart disease, most include extensive characterization of biologic traits and environmental characteristics, and many could be broadened to include additional phenotypes. Building upon these cohorts could leverage the ongoing investment in funding, infrastructure, and longitudinal data for the purposes of a large U.S. cohort study. Experienced investigators in existing cohort studies have demonstrated that they can collect high-quality data on a population basis, and many have established ongoing relationships with their communities. Response rates might also be higher among participants already committed to an ongoing research endeavor.

These advantages of existing cohorts are countered by concern for excessive burden on their participants and investigators, need for standardized data collection across sites and over time, desire for current state-of-the-art data collection methodology, limitations on data access and informed consent, complex administrative procedures for analysis and publication of data, and potential for important biases or lack of representativeness. Most of these cohorts, for example, involve middle-aged and older persons, while half the 2000 U.S. Census population was below age 35. A cohort sampled proportionately from these existing cohorts would have a very

different age distribution from the U.S. Census (Figure 1) and would under-represent Hispanics, African-Americans, and American Indians/Alaska Natives, as well as persons living in the West and having less than a college or high school education (Appendix 2).

More importantly, attempting to combine existing protocols and data, collected at different times for different purposes and with different methods, and rarely involving standardized ascertainment of more than a few disease outcomes, would make inferences drawn from such pooled data prone to important and irremediable biases related to the designs of the original studies. Selection of persons to be studied and methods that might be used in a potential large U.S. cohort study should be based on the most up-to-date scientific needs rather than on convenience or on past, potentially obsolete, study questions. A study of the size and scope needed to ascertain and quantify all the major environmental and genetic causes of common illnesses should be designed carefully and coherently from the start. Thus, although a potential new effort could seek to build on existing cohorts (see Section IV.C), new data collection and expanded consent would be needed from all study participants, and a new study design and infrastructure would be necessary.

III. PUBLIC CONSULTATION AND PARTICIPATION

III.A. Confidentiality and Participant Protections

Participants in a potential large U.S. cohort study of genes and environment would desire and deserve the most stringent confidentiality protections possible within the constraints of rigorous scientific design and balanced against the need for broad data access. Two over-riding principles should be followed: 1) data should not be placed in the public domain in formats that allow identification of individual participants; and 2) consent procedures should clearly define the scope and limitations of confidentiality protections.

III.B. Informed Consent

It would be optimal for such a study to utilize an open-ended consent, in which participants agree to have data incorporated into a large encrypted database that would serve as a resource for ongoing research, without that research being precisely specified in advance. Some communities might find open-ended consent less acceptable than others; the magnitude and impact of this concern could be examined in pilot studies and other public consultation activities, as outlined below. Because such a study would involve ongoing participation over a number of years, the need for periodic re-consenting should receive explicit consideration.

III.C. Central Institutional Review Board (IRB)

A central IRB oversight model would be ideal, to provide consistent study procedures and research oversight at all study sites. Although many institutions are resistant to non-local IRB oversight, in large part because they still retain responsibility for human subjects protection at their sites, precedents exist for successful central IRB strategies.

III.D. Need for Ongoing Public Consultation in Study Design and Implementation

Public consultation and pilot activities should be employed to help identify and address community concerns regarding population-based genetic research and explore the benefits or incentives that might be most important to different communities. The relative importance of individual incentives, inclusion of research topics of interest to the community, education for the community, etc. should be closely examined, as should the acceptability of confidentiality and privacy protection strategies.

Four types of public consultation activities should be considered:

- Information gathering: Surveys and focus groups should be used to inform study design and implementation within the limits that the proposed scientific requirements demand. They could identify preconceptions about this type of research, define expectations about privacy protections, obtain input on recruitment strategies and approaches for tailoring recruitment and consent processes, and develop educational and recruitment materials. Participants in both the surveys and focus groups should reflect the broad demographic makeup of the planned study cohort.

- Open dialogue: Town meeting-type fora would provide opportunities for dialogue and airing of concerns. They could begin an ongoing effort to capture community concerns and allow communities to identify research questions of particular interest. Town meetings should be geographically based, rather than stratified by individual population groups. All communities within a catchment area should be welcomed, although some might need to receive special invitation.

- Pilot investigations: The proposed recruitment strategy should be piloted to assess response rates, identify optimal recruitment and consent procedures, and assess alternatives. Multiple pilots in diverse regions and sociodemographic groups should be used to test recruitment procedures, determine response rates, and provide information about potential selection biases.

- Ongoing public consultation. Local and national media stories and print and on-line newsletters that discuss the study, its findings, and other health promotion issues would facilitate ongoing involvement of individual participants and their communities. Each study center should have mechanisms to assure ongoing involvement of their study communities and ongoing two-way flow of information between the communities and the centers. Such mechanisms might include community advisory boards, newsletters, town meetings, community-based health events, and local media stories.

IV. STUDY POPULATION

IV.A. Sample Size and Power Calculations

Large-scale population-based cohort studies are best suited for quantifying the risk associated with known genetic and environmental factors on a population basis, and for identifying genetic and environmental modifiers of this risk (gene-gene and gene-environment interactions). Assessment of interactions particularly requires very large sample sizes.

For qualitative traits (such as presence or absence of disease), sample size can be estimated by calculating the minimum number of cases, with two matched controls each needed to detect the desired increase in disease risk. While for some disorders, genetic or environmental risk factors with odds ratios of two or greater would be identified, for many common diseases the individual contributors would have odds ratios less than this. Even odds ratios of 1.2 to 1.5 can be important to detect and quantify, however, since they can provide critical clues to pathogenesis and potential intervention strategies. For example, the total risk to a sibling of a proband with type 2 diabetes (T2D) is 3.5, representing the sum of all of the genetic and environmental factors shared between sibs. Several individual genetic variants have now been identified for T2D, but none has an odds ratio greater than 1.5. Nonetheless, their significance remains profound; the identification of a predisposing variant in the PPARgamma gene, for example, has provided powerful validation of this orphan nuclear receptor for drug development. Indeed, the entire class of thiazolidinedione drugs for diabetes targets this receptor.

The number of cases needed to detect a range of gene (G), environment (E), G x G, and G x E effect sizes was calculated for a range of dominance models and allele frequencies for the risk allele and of exposure frequencies for the environmental factor. Population-based incidence estimates were used to determine the number of cases expected to accrue over five years in cohorts of 200,000, 500,000, and 1,000,000 participants with a 3% annual attrition rate (Table 2).

Figures 2 through 5 show the estimated minimum odds ratios for detecting main effects of G and E, and interactions of G x G and G x E, under a dominant genetic model with allele and environmental exposure frequencies of 10%. These odds ratios represent the minimum detectable effect with 80% power and type I error of 0.0001; power is greater to detect larger effect sizes or at higher type I error rates. For example, in Figure 2, all three sample sizes could detect odds ratios less than 1.5 for genetic main effects on the more common disorders. For rarer disorders, a cohort of 1,000,000 would be required, although 500,000 would detect odds ratios just over 2.0.

In general, power to detect genetic and environmental main effects would be acceptable for the two larger sample sizes for all frequencies examined. Power would also be good for G x G and G x E for all three cohort sizes for the more common diseases (incidence $\geq$ 200/100,000 per year). For less common disorders, a cohort size of 500,000 would give good power, but for rare diseases ($\leq$ 10/100,000 incidence per year) a cohort of 1,000,000 or more would be required.

Power would be lower to detect interaction effects in any subset of the cohort, such as age, ethnic, or geographic subgroups. In such subgroups, interactions specified as hypotheses stemming from other research would be tested based on prior evidence, so a higher type I error

rate is justified; power would thus increase. The cohort would then provide a "rapid response" sample for testing interaction hypotheses in a wide variety of subgroups for many common disorders.

For quantitative traits measured on a continuum, such as body mass index or visual acuity, all sampled individuals could be included in analyses. Sample size for quantitative traits can be determined using two gene-environment models, assuming that covariates account for 50% or 5% of the non-genetic variation and interact multiplicatively with the single locus effect. Sample sizes of 250,000, 100,000, 10,000, and 1,000 households were evaluated. Of note, for quantitative traits the analysis unit is a "household," assuming at least two unrelated individuals for analysis of variance and a trio (two parents and a child) for regression of offspring on mid-parent values. The number of sampled individuals would thus be two to three times the number of households. For virtually all models, the sample sizes evaluated would provide adequate power at significance levels of 0.01 and 0.0001. A genetically and clinically homogeneous sample of 10,000 households, consistent with the sizes of several key geographic or ethnic subgroups, would be adequate to detect single locus associations and gene-environment interactions when the variance of the individual effects is $\geq$1%. A sample of 5,000 households would detect these associations and interactions when the variance of individual effects is $\geq$ 5%. Heterogeneity of households would reduce study power and increase the sample size needed to detect a given effect.

Power would thus be preserved for quantitative traits if subgroups were examined, providing valuable insights into age-, race-, or geographic differences in genetic and environmental effects on important continuous biologic characteristics. Replication of findings by dividing the cohort into test and validation subsets would also be possible. With 250,000 participants, assuming 10% genetic heterogeneity, power would be 90% to detect a locus-specific heritability of 0.01. This level includes almost all of the continuous variables of current interest, such as body mass index, blood pressure, visual acuity, fasting glucose, total nevus count, cholesterol, and cognitive function.

IV.B. Representativeness and Yield

For such a potential study, the study cohort should be selected to match the most recent decennial U.S. Census population on six stratifying variables: age (five strata), sex (two strata), race/ethnicity (six strata), geographic region (four strata), education (four strata), and urban/rural residence (two strata) (Table 3).

Representativeness: The cohort should be recruited so that the final proportions in each of these six stratifying variables matches the most recent decennial census population. No attempt should be made to balance exactly across cross-stratifications, such as age-by-sex-by-race-by-region, since there are over 1,500 cross-strata. Data should be monitored during recruitment for important imbalances within key cross-strata, such as age-by-race or education-by-race. Sampling and recruitment strategies should be modified as needed to correct any major imbalances detected.

Yield: Every attempt should be made to maximize the response rate, but a new cohort study could likely do little better in this regard than other current large-scale cohort studies. Recent

experience suggests that, at best, only 40-60% of those contacted would agree to participate. Yield is likely to be lower among subgroups that have traditionally been less involved in research. This might require contacting more individuals in certain strata to achieve sufficient numbers in these subgroups. The genetic emphasis of the study might further complicate enrollment, as would a requirement (below) for full consent to sharing of data and samples. Yield and retention might be enhanced by the inclusion of family members, as described below.

IV.C. Sampling and Recruitment of Cohort Members

Participants should be sampled from defined census tracts as proposed by Field Center offerors responding to a Request for Proposals. The primary sampling unit should be the household, defined as any dwelling and all those living within it. An established sampling frame within each Field Center catchment area would be needed to select a representative sample of households and determine non-response rates. After a brief household enumeration, all members of the sampled household, whether biologically related or not, should be scheduled for a clinic visit. Participation of other household members should not be required for individual participation. Informed consent should be obtained at the clinic visit for each individual (not on a household basis), with appropriate procedures for assent for children under age 18. The final study cohort should include only participants completing the clinic visit and providing biologic samples. The proportion of biologically related individuals included in any Field Center should be limited to 30%; that is, at least 70% of the cohort should be unrelated individuals.

Inclusion Criteria: All persons residing in the sampled household who:

- are able to respond in the language of the interviewer;
- give informed consent for baseline examination, collection of DNA and white cells for possible transformation, and sharing of data and samples with IRB-approved investigators not directly affiliated with the study; and
- complete the baseline clinic examination and provide blood and other biologic samples.

Exclusion Criteria: Persons not meeting the inclusion criteria or who are:

- first degree relatives of a recruited participant at any Field Center that has exceeded its 30% cap on related individuals, to be monitored on a monthly- to quarterly-basis;
- members of existing cohorts in any age or race/ethnicity cell that contains 50% existing cohort members (see below);
- institutionalized persons (i.e., those in prison or long-term mental health or custodial care; persons in nursing homes or assisted living facilities, however, should be included); or
- in active military service.

Recruitment should primarily be door-to-door, although health maintenance organizations able to provide a population-representative cohort with defined census tract information should be eligible to serve as Field Centers and should not be expected to conduct door-to-door recruitment. Other creative and cost-effective approaches should be considered during protocol development.

Use of Existing Cohorts: Since information regarding population-based cohorts and their possible relationships with a potential large U.S. cohort study was important to the formulation of the present draft concept, a Request For Information (RFI) was issued to gather such information. As noted earlier, NIH supports a large number of population-based U.S. cohorts. Investigators in these cohorts were directly apprised of the RFI and a number expressed interest in participating in a large U.S. cohort study, were it to be undertaken. They were contacted again and invited to provide detailed information on their cohorts. Important demographic differences between these cohorts and the 2000 U.S. Census are detailed in Appendix 2. Given the wealth of experience and information available in these existing cohorts, investigators in such studies should be encouraged to apply and cohort participants with known census tract information should be invited to participate. The proportion of participants from existing cohorts should be limited to perhaps 50% of any age or race/ethnicity stratum, but this limitation should apply study-wide, not per Field Center. For instance, a single Field Center could recruit up to 3,000 children age four and under, or up to 2,000 Asian/Pacific Islanders, from existing cohorts, although care would be needed to avoid imbalance in other key stratifying variables.

Replacement of Cohort Attrition: A potential large U.S. cohort study should place a heavy emphasis on retaining participants who have completed the baseline examination, but a sizeable proportion would die, withdraw, or otherwise be lost in a ten-year period. In addition, despite the proposed recruitment of children born to study participants, the cohort would quickly age out of the zero to ten year age group. Even at an optimistic rate of loss of 3% of the cohort per year, after ten years only 76% would remain. If such a study were launched and were to continue beyond ten years, consideration should be given to recruiting a new subcohort designed either to match the cohort composition at entry or to meet the study goals or match U.S. demographics as they have evolved. For planning purposes, recruitment of a new subcohort of 25% of the original cohort size could be anticipated in years ten to 13.

V. DATA COLLECTION

V.A. Overview and Timeline

As described above, a range of sample sizes was considered: 200,000, 500,000, and 1,000,000 participants. For simplicity of exposition, the text describes a sample size of 500,000.

A 500,000 member cohort could be recruited and examined over a four-year period, following an initial year of public consultation, protocol development, requisite OMB approval, and training (Figure 6). A second exam could immediately follow the first and would also take four years to examine all participants. If the study were to continue, two more exams, each occurring over four years, could follow in sequence. If, instead, the study ended after ten years, a year of data analysis and close-out could follow the second exam. Some subgroups, such as the very young, the very old, or those with incident disease, might need to be seen more frequently. Participants should be contacted every six months (alternately by telephone and by mail or e-mail) to allow updating of contact information and ascertainment of incident outcomes. Disease outcomes

should be assessed using hospital records, outpatient records, and other data sources such as CMS data, registries, etc. Outcome ascertainment and data analysis could begin shortly after initiation of the first exam.

V.B. Baseline Examination

Detailed data collection at baseline would be vital to characterize participants' developmental and functional status, assess their current health, and identify cases of prevalent disease. Data collection at entry should include the widest breadth of phenotypes and environmental factors needed to predict outcomes, but cost and participant burden would be major concerns. Exam components should thus be prioritized and more efficient technologies developed on an ongoing basis. Such a study could use existing, validated data collection methods for the first exam, but should invest in development of innovative methodology for later exams (see Section VII).

A core group of baseline variables should be collected in all or nearly all participants, with additional variables added to the core list for different age groups. The final list of exam components should be determined during protocol development, after what is likely to be a lively debate, and should be based on the most up-to-date information and exemplar hypotheses. When designing the protocol, consolidating exam components (questionnaires, physical measurements, etc.) where possible would make optimal use of participant time. In addition, the use of mailed and Internet questionnaires should be encouraged. Table 4 provides a suggested list of exam components by age and their approximate duration. Maximum exam duration of four hours would be anticipated for a 500,000 member cohort.

Biologic specimens, including DNA and cells for possible transformation, should be collected from all study participants; potential participants refusing consent for collection or sharing of specimens should not be included in the cohort (see Inclusion Criteria above). Table 5 lists the types of specimens that should be collected and recommended number/size of aliquots. Measures that should be performed on all or nearly all participants are listed in Table 6, and should include the most extensive genotyping possible at the time the cohort is collected, with extension to complete DNA sequencing as the technology allows. The vast majority of specimens would be expected to be used in case-control or case-cohort analyses once sufficient incident cases accrue. Specimens should be stored at a contract-supported central repository with specified procedures for sample access and distribution (see Section VIII). A central CLIA-approved laboratory should be selected to develop biologic specimen collection protocols and perform study-wide analyses.

V.C. Follow-up Contacts and Examinations

Cohort follow-up would serve four purposes:

- identify outcomes of interest, including incident disease and developmental milestones;
- measure novel and state-of-the-art phenotypes and environmental exposures;
- assess change in phenotypes and environmental exposures; and
- establish a more stable baseline by repeated measures, where needed.

The cohort should be re-examined on average every four years. Follow-up examinations should reiterate the baseline exam in large part and use comparable methods as much as possible. Some baseline components should be deleted or postponed to allow time for more innovative measures as the science advances. The cohort should be recontacted by telephone annually, to alternate with annual mail- or web-based methods; participants would thus be contacted at least every six months.

V.D. Endpoints

A potential large U.S. cohort study would be intended to identify genetic and environmental contributors to major causes of human disease, disability, and premature morbidity as described in Table 1. Some diagnoses might be made as a consequence of examinations during a study clinic visit, but most could be ascertained through hospitalizations and physician office records, and by physician diagnoses reported by participants at the six-month contacts. While many such diagnoses might be reliable enough to accept without further investigation, review of records or specimens and adjudication using standard criteria might be necessary for conditions such as malignancies, major depression, dementia, heart failure, and stroke that are prone to misdiagnosis or inadequate characterization in routine clinical care. Conditions that might require an in-person visit or extensive questioning or examination include personality disorders, atherosclerosis, osteoarthritis, and alcohol and drug use. Record linkage systems (CMS, cancer registries, birth defect registries, etc.) should be used to enhance efficiencies of the ascertainment and validation systems. The NIH Roadmap initiatives on "Reengineering the Clinical Research Enterprise," particularly the establishment of a National Clinical Research Corps, could be highly complementary to the goals of a large U.S. cohort study. Similarly, the HHS plan for development of an electronic medical record (National Health Information Infrastructure, NHII) could be an enormous step forward in conducting this type of research in the U.S. As this is not likely to be fully in place until 2014, cohort study investigators could work with the HHS staff developing the NHII to ensure compatibility of the study's follow-up data systems. A potential study might also serve as a valuable test bed for the research applications of the NIH Roadmap and the NHII.

V.E. Notification of and Referral for Clinical Findings

One potential benefit of study participation could be the performance of multiple tests of possible clinical relevance. Participants indicating a desire to receive test results would be provided with those data, and, if the participant agreed, their physician would also be notified of findings needing medical follow-up. Less crucial information might be of interest to the participant and might promote continued participation. An initial report should summarize results available at the completion of each visit, such as blood pressure, visual acuity, and preliminary ECG findings. A second report should be sent as soon as routine laboratory and other centrally-determined results become available. Participants and, with their permission, their physicians should be immediately notified if potentially serious medical problems, such as severely elevated blood pressure, serious ECG abnormalities, or dangerously abnormal laboratory results were identified. A referral system

should be established based on the urgency of the need for medical attention, but might include conditions identified in Table 7.

Notification of genetic results is more challenging, due to concerns about potential genetic discrimination, limited availability of effective interventions, implications for family members, and other ethical, legal, and social issues. A potential study would retain its obligation to report research findings of definitive clinical value. Although some genetic tests have considerable potential for risk assessment and targeting of preventive strategies, those done in research settings today often do not predict the development and severity of complex diseases. This is expected to change, however, partly due to studies of the type described here. Reporting and counseling should be provided regarding genetic results clearly demonstrated to carry important health and management implications. On a regular basis, the study leadership should update the list of variants requiring reporting and counseling.

VI. INFORMATICS INFRASTRUCTURE AND DATA MANAGEMENT

Informatics and data management needs for a potential large U.S. cohort study would be formidable and include: 1) data capture, entry, and editing; 2) database design and management; and 3) analysis. The volume of data would require direct computer data entry or utilization of current technology for scanning of questionnaire data and computer touch screen queries. Further development of new technology for data collection would also be anticipated (see Section VII). As above, the projected implementation of electronic medical records could be of immense value to such a potential effort.

The design of the study database would be critically important for the successful development of the resource aspects of the study. Industry and academic experts have assured the expert panel that adequate computer hardware and software exist and can readily deal with a database of this size and complexity. All stressed, however, the importance of the design features that would drive such key elements as data integrity, ease of access, quality control, and analyses.

Exploratory analyses of the study cohort could be expected to begin as soon as the first wave of data is available. Further, specific hypotheses should be proposed that only a sample of this size could address, and new methods developed to minimize concerns regarding multiple hypothesis testing and spurious associations. Policies for accessing the data should facilitate both standard approaches and the development of new analytic techniques. A centralized resource for data analysis might be needed to insure full utilization of this resource.

VII. TECHNOLOGY DEVELOPMENT FOR DATA COLLECTION AND ANALYSIS

A potential large U.S. cohort study should devote at least 5% of its overall budget to technology development. Just as the Human Genome Project transformed our ability to collect and analyze genotypic data, such a study could provide a unique and historic opportunity to transform our

ability to collect and analyze both phenotypic and environmental exposure data. The value to biomedical research, health care, and biotechnology development of such progress in these two areas cannot be overemphasized.

VII.A. Development of Improved Technology to Collect and Analyze Phenotypic Data

New technologies are urgently needed for collection and analysis of phenotypic data. Ideally, they would be non-intrusive, applicable to all ages and health statuses, pain and risk free, reliable, inexpensive, self-calibrating, and able to transmit data linked to individual subjects directly to databases. Prototypes include real-time monitoring developed by NASA and DoD. A potential large U.S. cohort study would further their application to clinical research and routine health care. Table 8 lists technologies ripe for development and implementation within the next four years, and in five to ten years.

In coordination with ongoing programs at NHLBI, NIDDK, NIBIB, NIEHS, other ICs, DoD and NASA, technologies for both research and clinical applications could be developed that:

- Measure diet and nutritional status directly: Innovations in direct measurement, as by non-intrusive sampling of saliva or via an electronic record of intake, could provide more accurate and useful data than current food diary-based approaches.
- Monitor activity and physiological parameters: Personal sensors coupled via cell phone transmission to a data hub could achieve real-time, remote monitoring of these.
- Detect early disease: Examples include devices to measure bone density as an indicator of osteoporosis risk or to measure arterial flow as an indicator of atherosclerosis.
- Measure other biomarkers: Novel, high-content approaches for measuring an expanding catalog of proteomic and metabolomic markers could provide broad insights into disease status, general physiological functioning, and response to environmental exposures.

VII.B. Development of Technology to Collect and Analyze Environmental Exposure Data

Advances are also needed in measuring macro-scale and personalized environmental exposures (Table 9). Efforts could center on developing and applying new technologies to: 1) build a national exposure map; and 2) enable personalized exposure monitoring. A national exposure map could provide a seamless electronic infrastructure to identify and integrate existing environmental data sets for priority exposures, including toxicants and infectious agents. Many current data sources provide only broad geographical scale. A potential large U.S. cohort study could stimulate new technologies, such as macro- and micro-scale sensing devices for monitoring household and workplace exposures, to provide exposure data at finer geographical levels. New approaches to geographical information technology could be developed to link the data coverages spatially and temporally.

Technology development could overcome a major current impediment to population-based studies: lack of accurate, quantitative measures of exposure in relation to points of contact, internalization by the body, and early biological responses. In coordination with other NIH programs, other

interested federal agencies, and the private sector, new technologies and approaches could be developed to overcome such current obstacles in personal exposure monitoring as:

- Body Burden Assessment: Present body burden assays have only moderate sample throughput, detect a limited range of compounds, often are invasive and require clinical supervision, and are difficult to relate to biologically effective dose and to early biological responses, both of which link strongly to health outcomes.
- Exposure: Inability to determine physical activity that influences likelihood of exposure, to measure frequency and duration of exposure, and to assess the uptake and distribution in the body of environmental agents.
- Molecular Profiling: Inability to understand the biological processes that link environmental exposures to disease, including changes in the expression, activity, and interactions of genes, proteins, and metabolites.

VIII. ORGANIZATION

VIII.A. Participating Organizations

A potential cohort study of 500,000 individuals should include a Coordinating Center, roughly 100 Field Centers, a Central Laboratory/Repository, and an NIH Project Office. The Coordinating Center might subcontract for roughly five regional "hubs" to work more intensively on protocol adherence and quality control with groups of about 20 centers each.

VIII.B. Management

Management and monitoring of a large, complex cohort study would require a number of internal and external committees, as well as an established NIH Program Office and dedicated staff. Key internal committees are described in Table 10. Key external committees could include IC Representatives, Specimen Access, and Observational Study Monitoring Board:

- Committee of IC Representatives: ICs with an interest in the potential of such a study have named representatives to a preliminary version of such a Committee, which met in September 2004. If a large cohort study were to be undertaken, such a group should meet approximately monthly during protocol development and quarterly thereafter.

- Specimen Access Committee: A committee composed of an equal number of study investigators and unaffiliated scientists and an independent chair should be appointed to control access to specimens. Study investigators and outside investigators should follow the same procedures for requesting specimens, including obtaining IRB approval and signing a confidentiality agreement.

- Observational Study Monitoring Board (OSMB): The Director, NIH or his/her designee should establish an OSMB of 15 to 20 outside experts. The OSMB should monitor the data

from the study regularly, assess its progress, and make recommendations to the NIH on its conduct.

NIH management of the study should be through a Program Office comprised of roughly five professional level scientific staff, four support staff, and 12 to 15 contracting staff. Options for locating this Program Office include the NIH Office of the Director, a large Institute/Center with extensive population study experience (such as NCI, NHLBI, or NIDDK), or, because of its relevant expertise, NHGRI. Possible roles of the Project and Contracting Officer(s) are described in Table 11.

VIII. C. Policies

Data Access: Such a study should be a resource for the scientific community, both for investigators participating in the study directly and those outside it, nationally and internationally. Access to study data and biologic materials should thus be free and open. Investigators participating in the study should also have adequate opportunity to analyze and publish the results of their work. To balance these needs, the study should make all data available to study investigators and those outside the study at the same time. The informed consent documents should make it clear that individual, de-identified (but not permanently anonymized) phenotypic and genetic data would be shared with IRB-approved investigators from outside the study, including investigators from for-profit companies, who complete a confidentiality agreement (see Inclusion Criteria, above).

Intellectual Property: It is anticipated that important discoveries about genetic and environmental factors in health and disease would arise from such a large U.S. cohort study, although in many instances such a resource would serve not to discover but to validate, quantitate, and assess interactions between genetic and environmental risk factors that had been initially suggested by case-control studies. These discoveries could have potential commercial value in the diagnostic and therapeutic arena, however, and would often arise from the combination of the powerful public data resource provided by such a study and the ingenuity of an investigator who uses the resource to discover an important correlation. Insisting that all findings enter the public domain is an appealing strategy, and meshes with the overall philosophy of NIH and the Human Genome Project about openness. But care would need to be taken not to deter private sector investigators from using the resource or providing disincentives to the development of downstream diagnostic and therapeutic products that the public needs. Three possible ways to handle the intellectual property (IP) issues are most obvious, although other alternatives would also be possible:

- The NIH could insist that all discoveries made using the cohort study resource be placed in the public domain – potentially enforcing this by filing a "Declaration of Exceptional Circumstances" to the Bayh-Dole act, and making such an outcome a condition of access to the study database by investigators.

- The NIH could cede any decisions about IP to the investigator using the resource – although it is not clear how the PTO would handle a "discovery" where all of the data were generated

by an NIH study and were publicly accessible.

- The NIH could allow investigators using the study data to apply for IP as they see fit, but insist that they follow the NIH Research Tool Guidelines, and make a formal research exemption as a condition of use of the resource.

The last option would seem the most desirable but, before any such study were implemented, further discussion on this complex issue would be needed, including some attempt to harmonize the policy with that of other international cohort studies.

IX. CONCLUSION

We are poised at an extraordinary moment in humanity's history. The tools supplied by the Human Genome Project and other recent biomedical advances make it possible for the first time to accept the challenge of understanding the complex interplay between specific genetic and environmental factors that lead to health and disease – and to use that understanding to create a new age in health and health care that includes powerful new approaches to the prevention, diagnosis, and treatment of disease.

However, we can only achieve this key understanding if we first create the appropriate biomedical resources. The resource described here would provide a critical pathway by which the American people could benefit fully from the promise created by the Human Genome Project to improve human health. It could also provide an important foundation by which American researchers and biotechnology and pharmaceutical concerns could maintain their leadership in the life sciences.

Table 1. Potential conditions to be ascertained in a U.S. cohort, based on top 20 causes of DALY loss, death, and hospitalizations, with available incidence and prevalence estimates.

DALY Loss Rank (1996) Women	Men	Deaths thousands (2001)	Hospital-izations thousands (2001)	Incidence (cases/ 100,000/ year)	Prevalence thousands	Cause (ICD-10 or ICD-9 code)
1	1	502	2,090			Ischemic Heart Disease (I20-I25)
				565	7,800	Myocardial Infarction
		57	995	195	5,000	Heart Failure (I50)
3	6	164	931			Cerebrovascular Disease (I60-I69)
				177	4,800	Stroke
		19				Essential hypertension and hypertensive renal disease (I10,I12)
				2,155	65,000	Essential hypertension
		14				Atherosclerosis (I70)
		15				Aortic Aneurysm and Dissection (I71)
			1,633			Psychoses (290-299), especially:
2	10				13,600	Unipolar Major Depression
17						Post Traumatic Stress Disorder (F43.1)
19						Panic Disorder (F41.0)
20				--	10,447	Bipolar Disorder (F31)
				12		Schizophrenia

DALY Loss Rank (1996) Women	Men	Deaths thousands (2001)	Hospital-izations thousands (2001)	Incidence (cases/ 100,000/ year)	Prevalence thousands	Cause (ICD-10 or ICD-9 code)
		554	1,212	473	9,600	Malignant Neoplasms (C00-C97), especially:
5	3	156	153	66		Lung, Tracheal or Bronchial Cancer (C33-C34)
6		42	107	75	2,044	Breast Cancer (C50)
14	19	57	157	52	422	Colorectal Cancer (C18-C21)
	20	31		173		Prostate Cancer (C61)
9	11	71	562	204	11,967	Diabetes Mellitus (E10-E14)
4	7			--	11,200	Chronic Obstructive Pulmonary Disease (J40-J47), especially:
		17	505	116		Bronchitis and Emphysema (J40-J43)
12	16	4	454	177	16,000	Asthma (J45-J46)
		102				Other Chronic Lower Respiratory Diseases (J44, J47)
16		62	1,300			Lower Respiratory Infection (Influenza and Pneumonia, J10-J18)
		17				Pneumonitis (J69)
		32	315			Septicemia (A40-A41)
18	4	14				HIV/AIDS (B20-B24)
		17		18	452	Parkinson’s Disease (G20-G21)

DALY Loss Rank (1996) Women	Men	Deaths thousands (2001)	Hospital-izations thousands (2001)	Incidence (cases/ 100,000/ year)	Prevalence thousands	Cause (ICD-10 or ICD-9 code)
		54		146	4,000	Alzheimer's Disease (G30)
8	15			--	2,910	Dementia and other Degenerative/ Hereditary CNS Disorders
	17	27				Cirrhosis of Liver and Chronic Liver Disease (K70, K73-K74)
		39				Nephritis, Nephrotic Syndrome, Nephrosis (N00-N07, N17-N19, N25-N27)
				34	436	Renal Failure
15	18	14				Conditions Arising during Perinatal Period (P00-P96)
13	14					Congenital Abnormalities
7	12		496			Osteoarthritis
			999			Fractures, all sites
			315	106	--	Hip Fracture
	8	20				Homicide and Violence (U01-U02, X85-Y09, Y87.1)
	9	31				Self-Inflicted Injuries (U03, X60-X84, Y87.0)
		102				Accidents (V01-X59, Y85-Y86)

DALY Loss Rank (1996) Women	Men	Deaths thousands (2001)	Hospital-izations thousands (2001)	Incidence (cases/ 100,000/ year)	Prevalence thousands	Cause (ICD-10 or ICD-9 code)
10	2					Road Traffic Accidents
11	5					Alcohol Use
	13					Drug Use

Table 2. Estimated number of new cases available after 5 years of follow-up for varying cohort size and disease incidence rates per year, assuming 3% attrition per year.

Incidence/ 100,000/yr	Disease	1,000,000	500,000	400,000	300,000	200,000
10	Parkinson's Disease, SLE, Schizophrenia	457	228	183	137	91
50	Colorectal Cancer, Renal Failure, RA	2,282	1,141	913	684	456
100	Breast Cancer, Hip Fracture	4,559	2,279	1,824	1,368	912
200	Diabetes, Stroke, Heart Failure	9,100	4,550	3,640	2,730	1,820
500	MI, All Cancers	22,618	11,309	9,047	6,785	4,524
3,000	Cataracts, Hypertension	129,289	64,644	51,715	38,787	25,858

SLE = Systemic Lupus Erythematosus; RA = Rheumatoid Arthritis; MI = Myocardial Infarction

Table 3. Percent and projected number of 500,000 person cohort in each stratum of major characteristics.

Characteristic	Percent	Number (thousands) in 500,000 Person Cohort
Age (years)		
< 5	6	30
5-19	22	110
20-44	37	185
45-64	22	110
65-84	11	55
≥ 85	2	10
Sex		
Women	51	255
Men	49	245
Race/Ethnicity		
Hispanic	13	60
White, Non-Hispanic	69	345
Black, Non-Hispanic	12	60
Native American, Non-Hispanic	1	5
Asian/Pacific Islander, Non-Hispanic	4	20
Other, Non-Hispanic	2	10
Region		
Northeast	19	95
Midwest	23	115
South	36	180
West	23	115
Education (324,000 persons age ≥ 25)		
< High School	20	63
High School graduate	29	93
Some College	21	68
College Graduate	31	100
Urban/Rural		
Urban ≥ 500 persons/square mile	79	395
Rural < 500 persons/square mile	21	105

Table 4. Study components by age and their estimated duration in minutes, with shorter durations anticipated in children. 1M = components done in all three sample sizes; 500K = components done in 500K and 200K sample sizes only; 200K = components done in 200K sample size only.

		Age Group (years)						
Study Component	**Duration**	**0-5**	**6-11**	**12-19**	**20-34**	**35-49**	**50-64**	**65+**
All or Nearly All Participants								
Household Enumeration	5			1M	1M	1M	1M	1M
Informed Consent	15-30	1M	1M	1M	1M	1M	1M	1M
Tracking Information (contact, next of kin, etc)	5			1M	1M	1M	1M	1M
Demographics (age, SES, etc)	5	1M	1M	1M	1M	1M	1M	1M
Medical History (self-report of doctor's diagnosis)	5-10	1M	1M	1M	1M	1M	1M	1M
Prescription Medications, OTC medications, vitamins/supplements	5-10	1M	1M	1M	1M	1M	1M	1M
Food Frequency/Dietary History (including eating patterns and food availability)	15	1M	1M	1M	1M	1M	1M	1M
Physical Activity (overall/leisure time/occupational)	10-15	1M	1M	1M	1M	1M	1M	1M
School or Work Absence	5		200K	200K	200K	200K	200K	200K
Family History	5-10	1M	1M	1M	1M	1M	1M	1M
Environmental Exposure Questionnaire	10-15	1M	1M	1M	1M	1M	1M	1M
Anthropometry (height, weight, waist/hip circumference)	10	1M	1M	1M	1M	1M	1M	1M
Heart Rate/Blood Pressure	10	1M	1M	1M	1M	1M	1M	1M
Visual Acuity Screening	5-10	500K	500K	500K	500K	500K	500K	500K
Hearing Acuity Screening	5-15	200K	200K	200K	200K	200K	200K	200K

		Age Group (years)						
Study Component	**Duration**	**0-5**	**6-11**	**12-19**	**20-34**	**35-49**	**50-64**	**65+**
Biologic Specimen Collection	15	1M	1M	1M	1M	1M	1M	1M
Sleep Habits	10	200K	200K	200K	200K	200K	200K	200K
Occupation History (current/usual)	5-10			500K	500K	500K	500K	500K
Medical Care Access/Use	5	500K	500K	500K	500K	200K	200K	200K
Oral Health/Dental Exam	15	200K	200K	200K	200K	200K	200K	200K
Age: 0-5 years								
Developmental milestones	20	1M						
Additional Medical History	5	1M						
Safety Questionnaire	5	500K						
Maternal History/Prenatal Exposures	10	1M						
Head Circumference	5	500K						
Age: 6-19 years								
Depression/Anxiety/ADHD/Conduct Disorders (ASPD)/Behaviors/Adverse Childhood Environment (ACE)	15		1M					
Measures of School Performance	10		1M					
Focused Medical History	5		1M					
Safety	5		500K					
Behavioral Traits/Peer Pressure	5		500K					
Eating disorders/TV/Computing time	5		1M					
Weight Fluctuation/Cycling	5		200K					

Study Component	Duration	Age Group (years) 0-5	6-11	12-19	20-34	35-49	50-64	65+
Age: 12 years and Older								
Smoking/Tobacco Use	5			1M	1M	1M	1M	1M
Alcohol/Drug Use/Abuse Dependency	10			1M	1M			
Reproductive History (women)	10			1M	1M	1M	1M	1M
Focused Medical History	5			1M	1M	1M	1M	1M
Health Related Quality of Life/SF36	10			500K	500K	500K	500K	500K
Use of Preventive Screening Measures	5			200K	200K	200K	200K	200K
Age: 35 years and Older								
Depression/Anxiety/Behaviors	15					1M	1M	1M
Self-Reported Health Status	1					500K	500K	500K
Caregiving/Parenting	5					200K	200K	200K
Age: 50 years and Older								
ECG	15						1M	1M
Cognitive Function Testing	10						1M	1M
Age: 65 years and Older								
Activities of Daily Living (ADL)	5							500K
Timed Walk\Chair Stand	10							1M
Additional Medical History	5							500K

		Age Group (years)						
Study Component	**Duration**	**0-5**	**6-11**	**12-19**	**20-34**	**35-49**	**50-64**	**65+**
Total Minutes, 500K cohort	[240]	160	165	195	220	216	231	246
Total Hours, 500K cohort	[4.0]	2.67	2.75	3.25	3.67	3.60	3.85	4.10
Total Minutes, 200K cohort	[300]	195	230	260	285	296	321	341
Total Hours, 200K cohort	[5.0]	3.25	3.83	4.33	4.75	4.93	5.35	5.68
Total Minutes, 1M cohort	[180]	140	140	165	185	180	195	200
Total Hours, 1M cohort	[3.0]	2.33	2.33	2.75	3.08	3.00	3.25	3.33

200K cohort: do everything in everyone
500K cohort: delete school/work absence, hearing acuity, sleep habits, oral health/dental exam, weight fluctuation/cycling, preventive screening, caregiving
1M cohort: also delete visual acuity, occupational history, medical care access, safety, head circumference, behavioral traits/peer pressure, SF36, self-reported health, ADL, additional medical history > 65

Table 5. Biologic specimens to be collected at baseline and subsequent exams.

Type	Timing of Collection	Size/Number of Aliquots
Serum/Plasma	Every Exam	500 uL/70 aliquots
White Cells/Buffy Coats/DNA	Every Exam	pellet from one 7ml tube/5 aliquots
Red Cells	Baseline	one serum separator tube
Spot Urine	Every Exam	15ml/2 aliquots
Nails/Hair	Baseline	
Saliva	Baseline	
Cancer Tumor Tissue Block	As Available in Follow-Up	
Placenta/Cord Blood	As Available in Follow-Up	
Exfoliated Teeth	As Available in Follow-Up	
Brain Bank	As Available in Follow-Up	

Table 6. Laboratory measures to be performed in all participants.

Measure	**Timing of Measure**
CBC	Baseline
Fasting Total/HDL Cholesterol and Triglycerides	Baseline and follow-up exams
Fasting Glucose	Baseline and follow-up exams
PSA (men > 35 only)	Baseline and follow-up exams
C-reactive Protein	Baseline and follow-up exams
Inflammatory Markers (Fibrinogen Antigen, Interleukin-6; > 35 only)	Baseline and follow-up exams
Renal Function Markers (Creatinine, Microalbuminuria)	Baseline and follow-up exams
Viral/Serologic Assays	Baseline and follow-up exams
Lead Exposure (ages 0-19 only)	Baseline and follow-up exams
Folate, Vitamin B12	Baseline and follow-up exams
Genotyping	
Association Studies on Candidate Genes	Baseline and ongoing
Whole Genome Association Studies with HapTag SNPS	As cost allows
Complete Genome Sequencing	As affordable technology becomes available

Table 7. Levels for notification and referral of abnormal exam findings.

Finding	**Alert Level**
Acute ST segment elevation, ventricular tachycardia	Immediate
Unstable angina	Immediate
Neurologic symptoms in past week	Immediate
Suicidal ideation	Immediate
Systolic BP $\geq$ 210	Immediate
Diastolic BP $\geq$ 120	Immediate
180 < Systolic BP $\leq$ 210	Urgent
110 < Diastolic BP $\leq$ 120	Urgent
Total cholesterol $\geq$ 360 mg/dL	Urgent
Triglyceride $\geq$ 1000 mg/dL	Urgent
HDL cholesterol $\leq$ 20 mg/dL	Urgent
Calculated LDL cholesterol $\geq$ 260 mg/dL	Urgent
Fasting glucose $\leq$ 50 mg/dL or $\geq$ 400 mg/dL	Urgent
Creatinine $\geq$ 2.0 mg/dL	Urgent
Elevated lead levels under age 5	Urgent

Immediate = participant is taken from field center directly to physician or hospital.
Urgent = participant and physician (if participant agrees) notified within one week of receipt of study results.

Table 8. Development of technologies for more accurate phenotypic assessment.

	First Generation Technologies 0-4 years	Second Generation Technologies 5-10 years
Measurement of Physiology	• Non-intrusive devices such as rings, bracelets, and body patches to measure heart rate and oxygen saturation simultaneously • Minimally intrusive sensors for glucose and lactate, which could be leveraged to measure many enzymatic products • Determine nutritional status from nail or hair clippings • Data analysis by third party software on phone	• "Smart" clothing • Methods and instrumentation to measure thousands of analytes • Improved use of respiratory products as a source of physiologic data • Proteomic and metabolomic profiling
Measurement of Activity	• Combination of accelerometers, HR monitors, and continuous monitoring to measure posture, ambulation, running, exercise, etc. • Use of cell-phones with built-in GPS as hubs for data collection (also applicable to measurement of physiology and diet)	• "Smart" house technologies

Table 9. Development of technologies for assessment of environmental exposures.

Level of Difficulty	**First Generation Technologies 0-4 years**	**Second Generation Technologies 5-10 years**
High ↑	• Develop methods to integrate data coverages electronically for a national exposure map.	• Develop multiplexed, macro-scale environmental sensors for known ambient and personal exposures. • Develop integrated sensor networks. • Develop deployable micro- and nano-scale biosensors. • Develop micro-scale environmental sensors for unknown exposure agents. • Make a national exposure map a publicly available resource.
Moderate ↑	• Develop GIS coverages with existing data sets. • Improve sample matrix selection and analysis for body burden assessment and molecular profiling. • Improve methods of sample preparation and analysis for molecular profiling and body burden assays. • Conduct molecular profiling studies in animals and human cell lines to guide in the interpretation of molecular profile data for this cohort.	• Develop baseline ranges of exposures for this cohort. • Integrate geographic information systems data to identify environmental risk factors for disease and at risk subpopulations. • Develop diagnostic biosensors for known exposure agents. • Initiate new studies to address data gaps for a national exposure map.
Low	• Select preferred technologies for molecular profiling. • Develop data and technology standards for transciptomics, proteomics, and metabolomics. • Select data sets for geographic information systems coverages. • Develop a list of priority environmental agents and environmental stressors. • Develop wearable personal biosensors for monitoring activity patterns.	• Expand geographic information systems coverages to incorporate data from this cohort. • Develop micro-scale environmental sensors for known exposure agents. • Develop new methods for assessing biologically effective doses for specific compounds.

Table 10. Possible study committees and charges.

Commitee	Charge
Steering	• Develop and modify study protocol and policies • Review Coordinating Center reports of study progress and recommend corrective action to NIH as needed • Review and approve recommendations of study subcommittees for approval by NIH
Executive	• Handle day-to-day study decision-making and management • Inform Steering Committee regularly of actions and decisions
Design	• Evaluate and prioritize proposed exam components and make recommendations to Steering Committee regarding exam content • Consider timing of components over course of study, repetition of component, participant burden, and cost, along with scientific value.
Recruitment	• Evaluate and recommend modifications to recruitment strategies to assure consistency of procedures among Field Centers, as feasible. • Evaluate status of recruitment, considering balance of relevant ethnic, gender, and age subgroups • Develop a standard description of recruitment procedures for use in study manuscripts
Operations	• Evaluate recommended exam components for participant burden; operationalize approved components • Make recommendations to Steering Committee regarding methods to minimize participant burden and optimize comfort, interest, and satisfaction • Assure participant concerns are addressed and ensure maximum participation and retention • Develop methods to train exam staff; plan and execute training for exam procedures; develop procedures for exam technicians to obtain and maintain certification to perform study procedures • Develop Manual of Operations for clinic operations. • Develop a regular newsletter to keep participants informed about study and foster good will • Develop system of "alert" values and procedures for providing feedback to and referrals for participants and their health care providers

Quality Control	• Develop methods to assess accuracy and reliability of exam methods and control variability, including collection of quality control data • Evaluate quality control data, report to Steering Committee regularly, alert Steering Committee when reliability or variability are unacceptable, and recommend and oversee further investigation and corrective action
Laboratory	• Recommend specimen-based laboratory measurements; develop protocol for Field Center specimen collection • Recommend plan for quality assurance, and develop and recommend methods to assess comparability among centers and to investigate reasons for lack of comparability or unacceptable variability among Field Centers or within a Field Center • Recommend further investigation and corrective action
Events	• Develop protocol for identifying, evaluating, and quantifying, as feasible and appropriate, study outcomes • Participate in classification of type and severity of outcomes
Ancillary Studies	• For studies to be funded from other than contract funds, review, recommend modifications to science and logistical conduct, and recommend approval or disapproval to Steering Committee
Publications	• Develop, disseminate, and enforce policies for proposing and conducting data analyses; establishing authorship and reinforcing responsibilities of authorship; monitoring progress of data analyses; and use of data in abstracts, presentations, and publications. • Develop and assist in maintenance of publications data base by the Coordinating Center • Recommend to Steering Committee directions for publications and presentations. • Review, recommend modifications for, and consider for approval all abstracts, presentations, manuscripts, and other data analyses emanating from the study
Access	• Develop study procedures for access to specimens and data • Monitor access procedures to ensure participant confidentiality, compliance with informed consent, and openness of access

Table 11. Roles of project officer(s) and contracting officer(s).

Project Officer	Contracting Officer
• Participate in the Steering Committee and its subcommittees in protocol development.	• Participate in the Steering Committee and its subcommittees to assure that study resources are used within funding allotments and in accordance with contractual requirements.
• Ensure the study meets its scientific objectives while remaining on schedule and within budget, and work with the Steering Committee to resolve any technical problems that arise.	• Provide the Project Officer an interpretation of contractual requirements.
• Monitor the progress of the study by maintaining close contact with investigators, inspecting and accepting contract deliverables, and performing periodic site visits.	• Monitor the study expenditures and deliverables, recommend appropriate action to the Project Officer and upon the Project Officer's approval authorize any required action.
• Assist the Contracting Officer in authorizing reimbursement of costs and in negotiating any changes in the contract Statements of Work, periods of performance, or delivery schedules.	• Assist the Project Officer in negotiating any funding and/or contractual changes and upon the Project Officer's approval authorize funding and/or contractual changes.
• Participate in analysis and publication of study results.	

Figure 1. Age distribution of the US Population, Census 2000, and of existing cohorts responding to Request for Information.

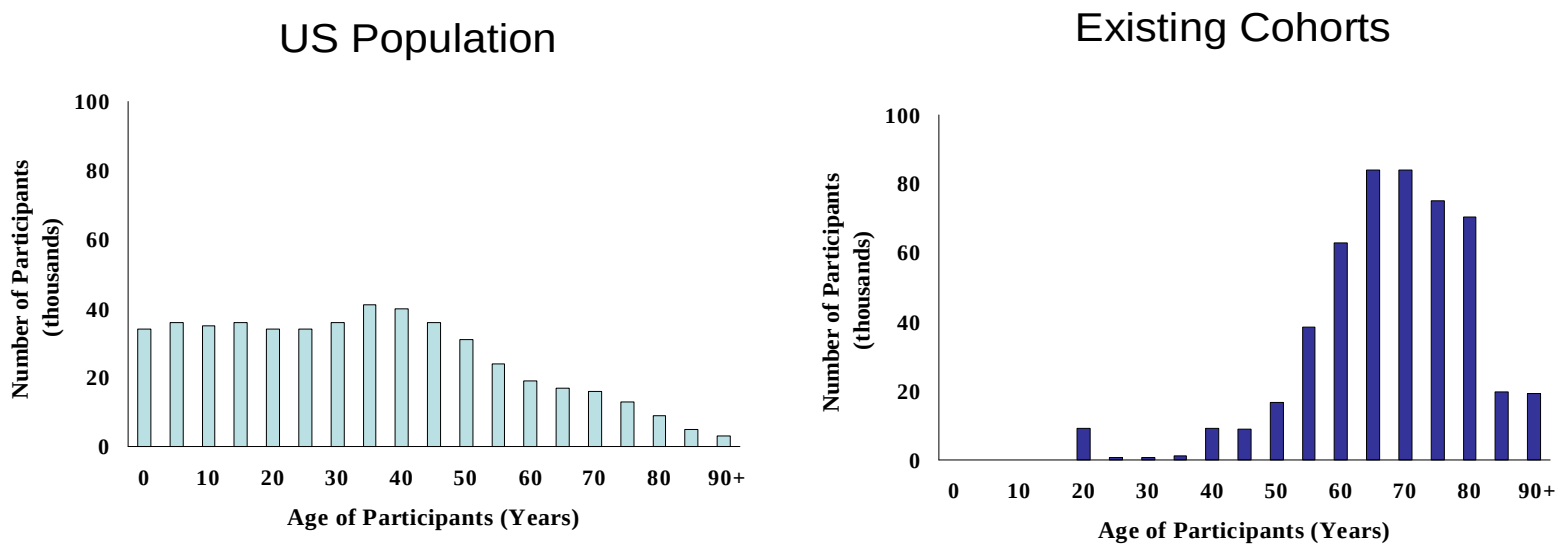


Figure 2.

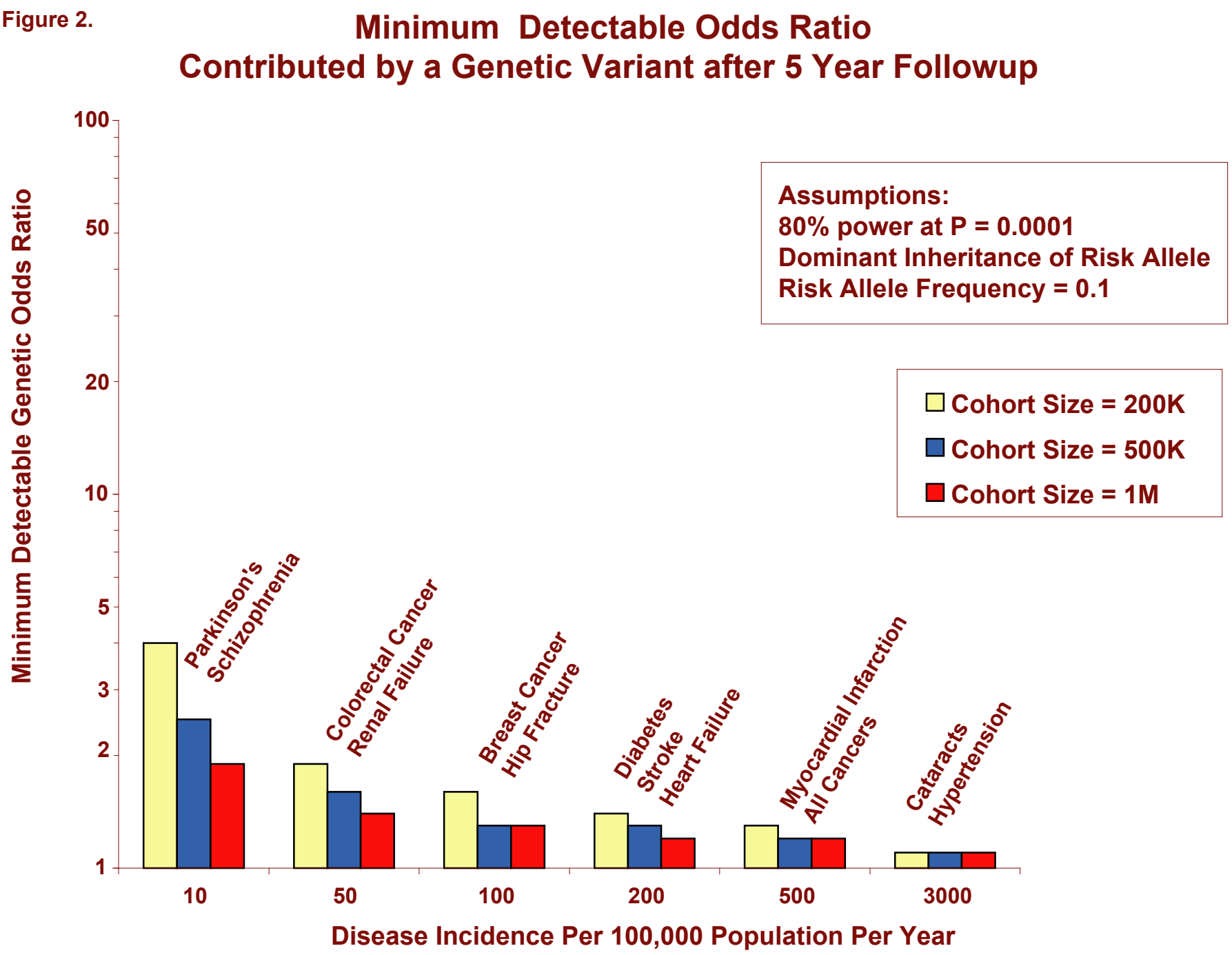

Minimum Detectable Odds Ratio
Contributed by a Genetic Variant after 5 Year Followup
Minimum Detectable Genetic Odds Ratio
100
50
20
10
5
3
2
1
Assumptions:
80% power at P = 0.0001
Dominant Inheritance of Risk Allele
Risk Allele Frequency = 0.1
Cohort Size = 200K
Cohort Size = 500K
Cohort Size = 1M
Parkinson's
Schizophrenia
Colorectal Cancer
Renal Failure
Breast Cancer
Hip Fracture
Diabetes
Stroke
Heart Failure
Myocardial Infarction
All Cancers
Cataracts
Hypertension
10
50
100
200
500
3000
Disease Incidence Per 100,000 Population Per Year

Figure 3.

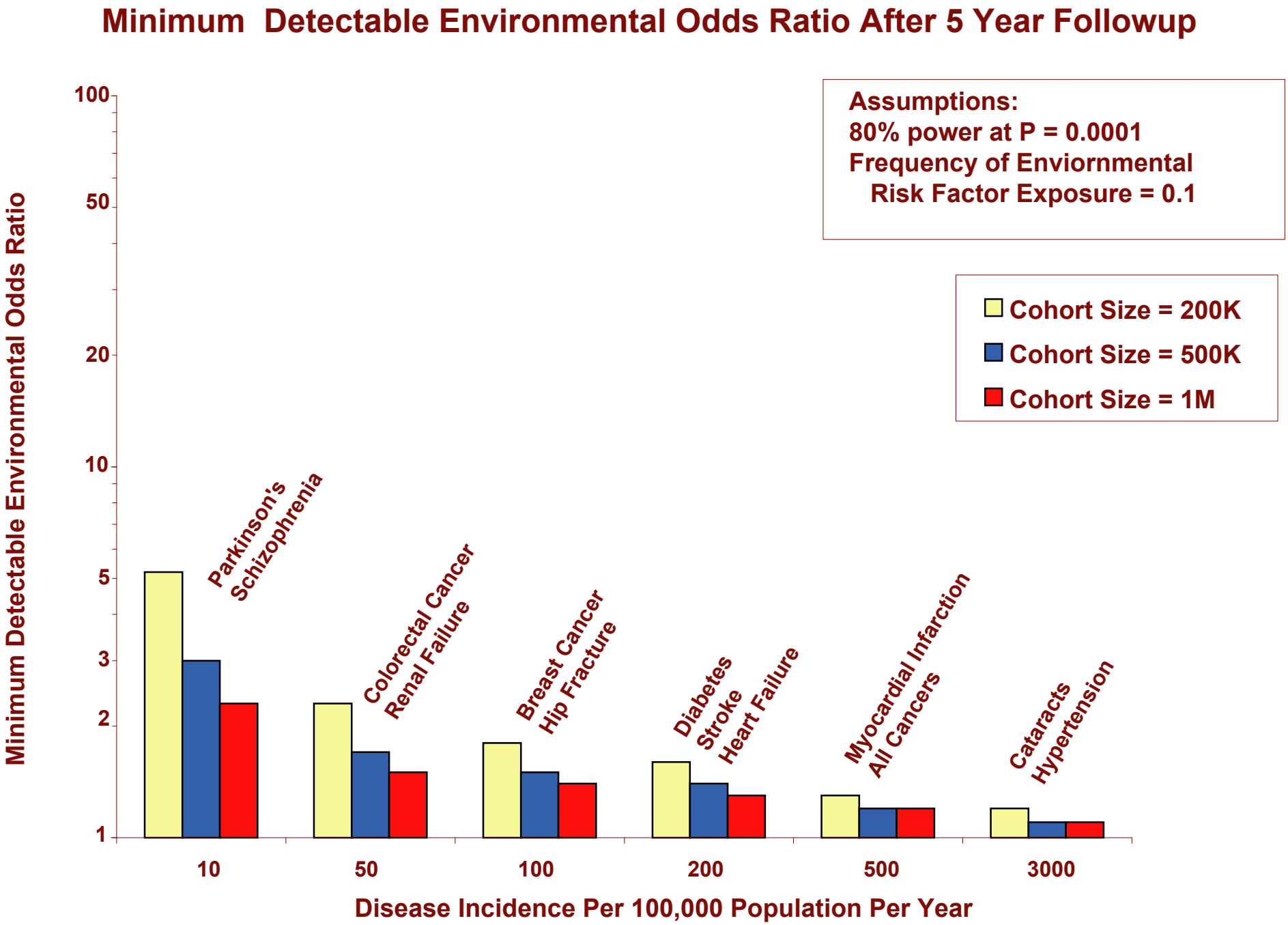

Minimum Detectable Environmental Odds Ratio After 5 Year Followup
Assumptions:
80% power at P = 0.0001
Frequency of Enviornmental
Risk Factor Exposure = 0.1
Cohort Size = 200K
Cohort Size = 500K
Cohort Size = 1M
Minimum Detectable Environmental Odds Ratio
100
50
20
10
5
3
2
1
Parkinson's
Schizophrenia
Colorectal Cancer
Renal Failure
Breast Cancer
Hip Fracture
Diabetes
Stroke
Heart Failure
Myocardial Infarction
All Cancers
Cataracts
Hypertension
10
50
100
200
500
3000
Disease Incidence Per 100,000 Population Per Year

Figure 4.

Minimum Detectable Gene-Gene (GxG) Interaction Odds Ratio After 5 Year Followup

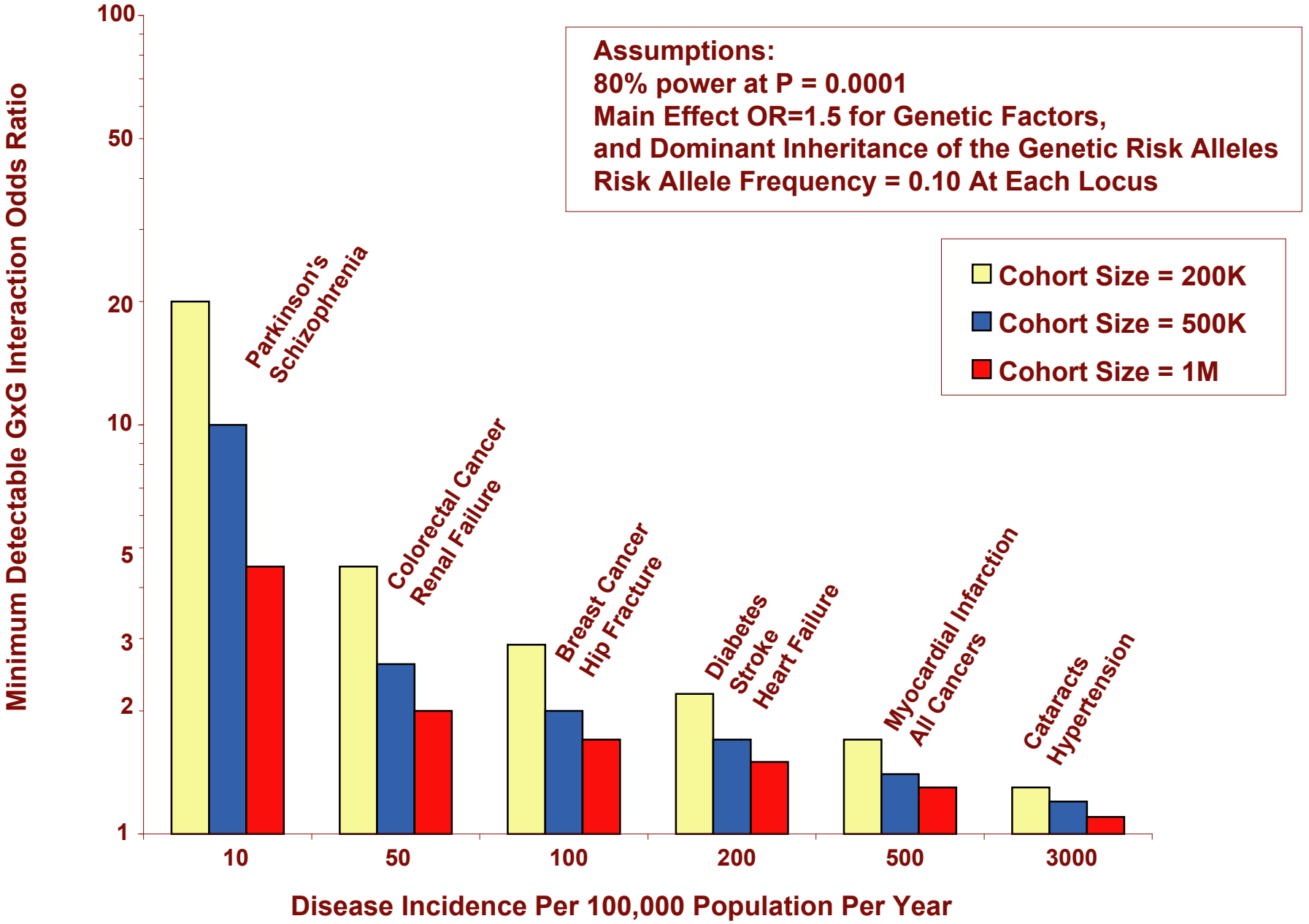


Figure 5.

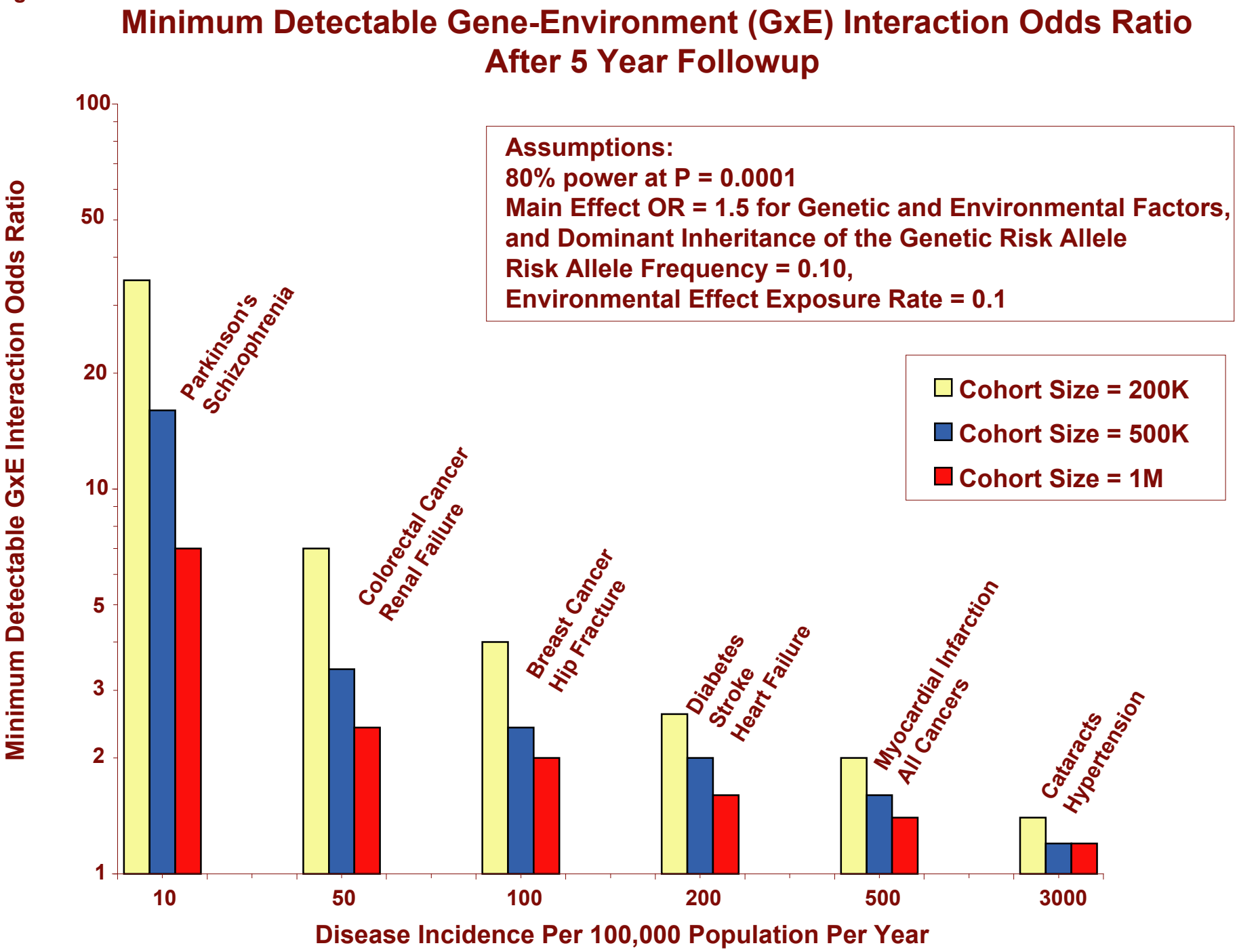
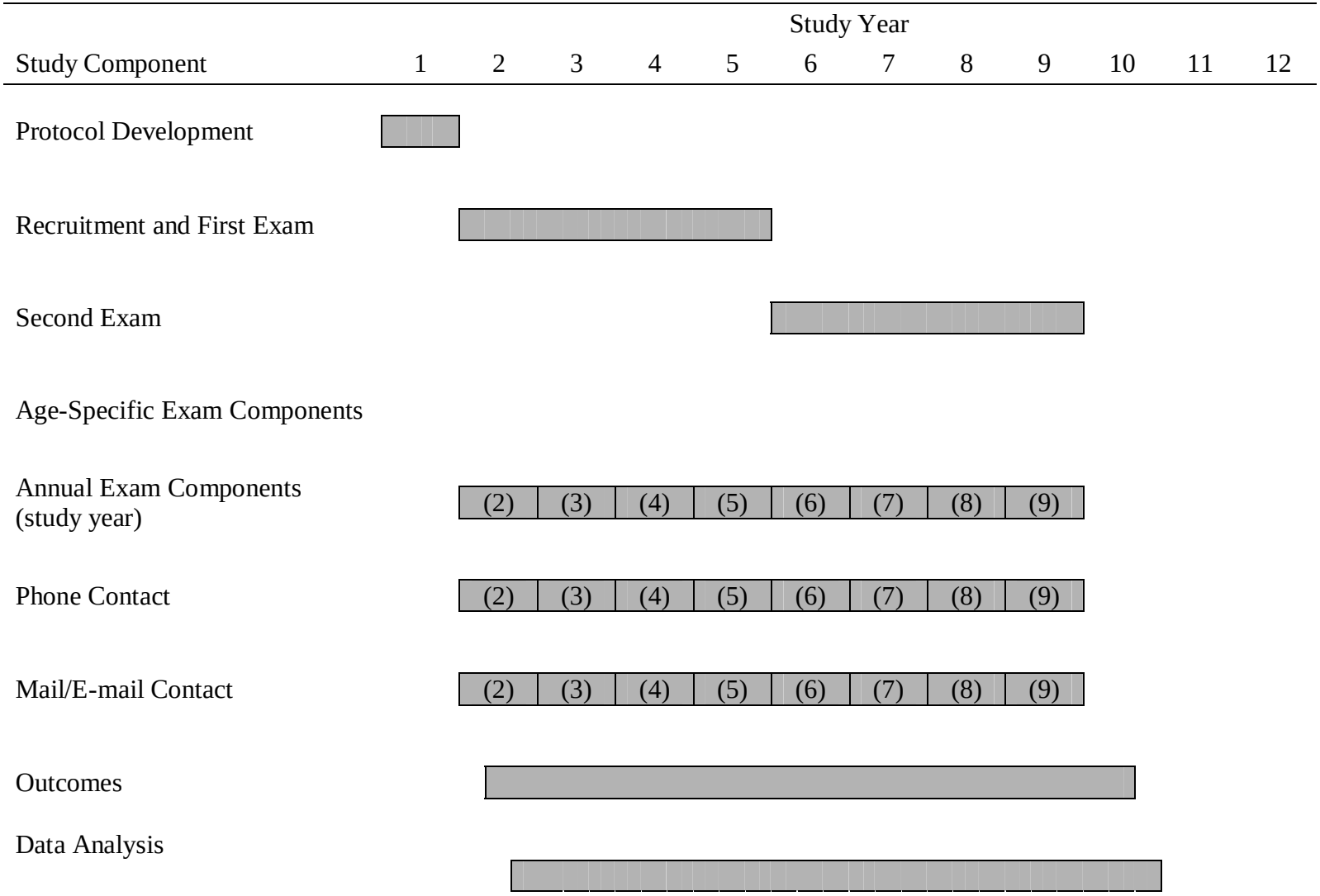

Minimum Detectable Gene-Environment (GxE) Interaction Odds Ratio
After 5 Year Followup
Assumptions:
80% power at P = 0.0001
Main Effect OR = 1.5 for Genetic and Environmental Factors,
and Dominant Inheritance of the Genetic Risk Allele
Risk Allele Frequency = 0.10,
Environmental Effect Exposure Rate = 0.1
Cohort Size = 200K
Cohort Size = 500K
Cohort Size = 1M
Minimum Detectable GxE Interaction Odds Ratio
100
50
20
10
5
3
2
1
Parkinson's
Schizophrenia
Colorectal Cancer
Renal Failure
Breast Cancer
Hip Fracture
Diabetes
Stroke
Heart Failure
Myocardial Infarction
All Cancers
Cataracts
Hypertension
10
50
100
200
500
3000
Disease Incidence Per 100,000 Population Per Year

Figure 6. Study Component Timeline for 500,000 Person Cohort.



Study Component	Study Year 1	2	3	4	5	6	7	8	9	10	11	12
IF STUDY IS CONTINUED												
Third Exam												
Recruitment of 25% "Replacement" Subcohort												
Annual Exam Components										(10)	(11)	(12)
Phone Contact										(10)	(11)	(12)
Mail/E-mail Contact										(10)	(11)	(12)
Outcomes												
Data Analysis												

APPENDIX 1

WORKING GROUP ROSTER

David Altshuler
Harvard Medical School
Massachusetts General Hospital

Joan E. Bailey-Wilson
National Human Genome Research Institute
National Institutes of Health

Eric Boerwinkle
University of Texas

Gregory L. Burke
Wake Forest University School of Medicine

Wylie Burke
University of Washington

C. Christopher Hook
Mayo Medical School

Rodney Howell
National Institute of Child Health and Human Development
National Institutes of Health

Jean MacCluer
Southwest Foundation for Biomedical Research

Donald R. Mattison
National Institute of Child Health and Development
National Institutes of Health

Jeffrey C. Murray
University of Iowa

Larry L. Needham
Centers for Disease Control and Prevention

M. Anne Spence
University of California, Irvine

Alexander F. Wilson
National Human Genome Research Institute
National Institutes of Health

Samuel H. Wilson
National Institute of Environmental Health Sciences
National Institutes of Health

Staff:

Francis Collins
National Human Genome Research Institute
National Institutes of Health

Alan E. Guttmacher
National Human Genome Research Institute
National Institutes of Health

Teri Manolio
National Heart, Lung, and Blood Institute
National Institutes of Health

Laura Lyman Rodriguez
National Human Genome Research Institute
National Institutes of Health

Susan Vasquez
National Human Genome Research Institute
National Institutes of Health

SUB-GROUP MEMBERSHIP

Data Collection Sub-Group:

Greg Burke
Wake Forest University

Virginia Cain
Office of Behavioral and Social Science Research
National Institutes of Health

Gwen Collman
National Institute of Environmental Health Sciences
National Institutes of Health

Bob Davis
Centers for Disease Control and Prevention

Bridget Grant
National Institute on Alcohol Abuse and Alcoholism
National Institutes of Health

Barbara Klein
University of Wisconsin

Teri Manolio
National Heart, Lung, and Blood Institute
National Institutes of Health

Kathleen Merikangas
National Institutes of Mental Health
National Institutes of Health

Vivian Ota Wang
National Human Genome Research Institute
National Institutes of Health

Eric Rimm
Harvard University

Ronald Ross
University of Southern California

Community Engagement Sub-Group:

Wylie Burke
University of Washington

Alan Guttmacher
National Human Genome Research Institute
National Institutes of Health

Jeff Long
University of Michigan

Colleen McBride
National Human Genome Research Institute
National Institutes of Health

Jean McEwen
National Human Genome Research Institute
National Institutes of Health

Jeff Murray
University of Iowa

Laura Rodriguez
National Human Genome Research Institute
National Institutes of Health

Charmaine Royal
Howard University

Phenotyping Technology Sub-Group:

David Altshuler
Harvard University

Steve Carr
Broad Institute, MIT

Geoff Duyk

Abby Ershow
National Heart, Lung and Blood Institute
National Institutes of Health

Stephen Intille
Massachusetts Institute of Technology

David Klonoff
Mills-Peninsula Health Services

Maren Laughlin
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health

Brad Ozenberger
National Human Genome Research Institute
National Institutes of Health

Allison Peck
National Human Genome Research Institute
National Institutes of Health

Lloyd Smith
University of Wisconsin

Sample Sub-Group:

Eric Boerwinkle
University of Texas

Gerardo Heiss
University of North Carolina

Kathy Helzlsouer
Johns Hopkins University

William Kalsbeek
University of North Carolina

Jill Norris
University of Colorado

Diana Petitti
Kaiser Permanente

Margaret Spitz
M.D. Anderson Cancer Center

Marcia Stefanick
Stanford University

Teri Manolio
National Heart, Lung, and Blood Institute
National Institutes of Health

Environmental Exposure Technology Development Sub-group

David Balshaw
National Institute of Environmental Health Sciences
National Institutes of Health

John Barr
Centers for Disease Control and Prevention

David Brown
National Institute of Environmental Health Sciences
National Institutes of Health

Mark Ellisman
University of California, San Diego

Alan Guttmacher
National Human Genome Research Institute
National Institutes of Health

Bruce Hammock
University of California, Davis

Paul Lioy
UMDNJ/Rutgers

Teri Manolio
National Heart, Lung, and Blood Institute
National Institutes of Health

Gilbert Omenn
University of Michigan

Martin Philbert
University of Michigan

John Potter
Fred Hutchinson

Leona Samson
Massachusetts Institute of Technology

Jeff Schloss
National Human Genome Research Institute
National Institutes of Health

Martyn Smith
University of California, Berkeley

Lydia Sohn
University of California, Berkeley

William Suk
National Institute of Environmental Health Sciences
National Institutes of Health

Susan Sumner
RTI

James Swenberg
University of North Carolina

Claudia Thompson
National Institute of Environmental Health Sciences
National Institutes of Health

Simon Watkins
University of Pittsburgh

David Walt
Tufts University

Brenda Weis
National Institute of Environmental Health Sciences
National Institutes of Health

Samuel Wilson
National Institute of Environmental Health Sciences
National Institutes of Health

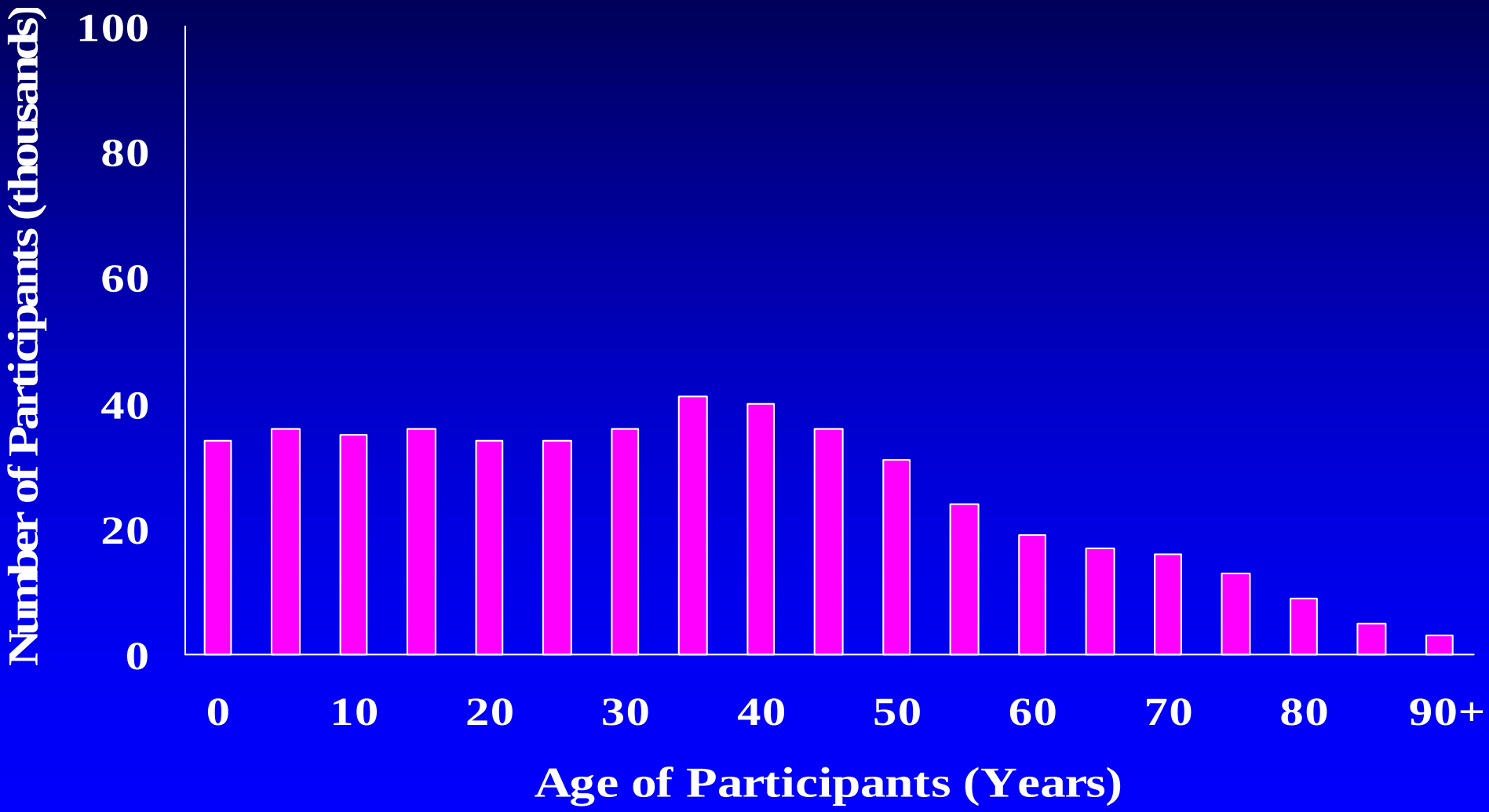
APPENDIX 2

REQUEST FOR INFORMATION: DESIGN AND IMPLEMENTATION OF A LARGE-SCALE PROSPECTIVE COHORT STUDY OF GENETIC AND ENVIRONMENTAL INFLUENCES ON COMMON DISEASES

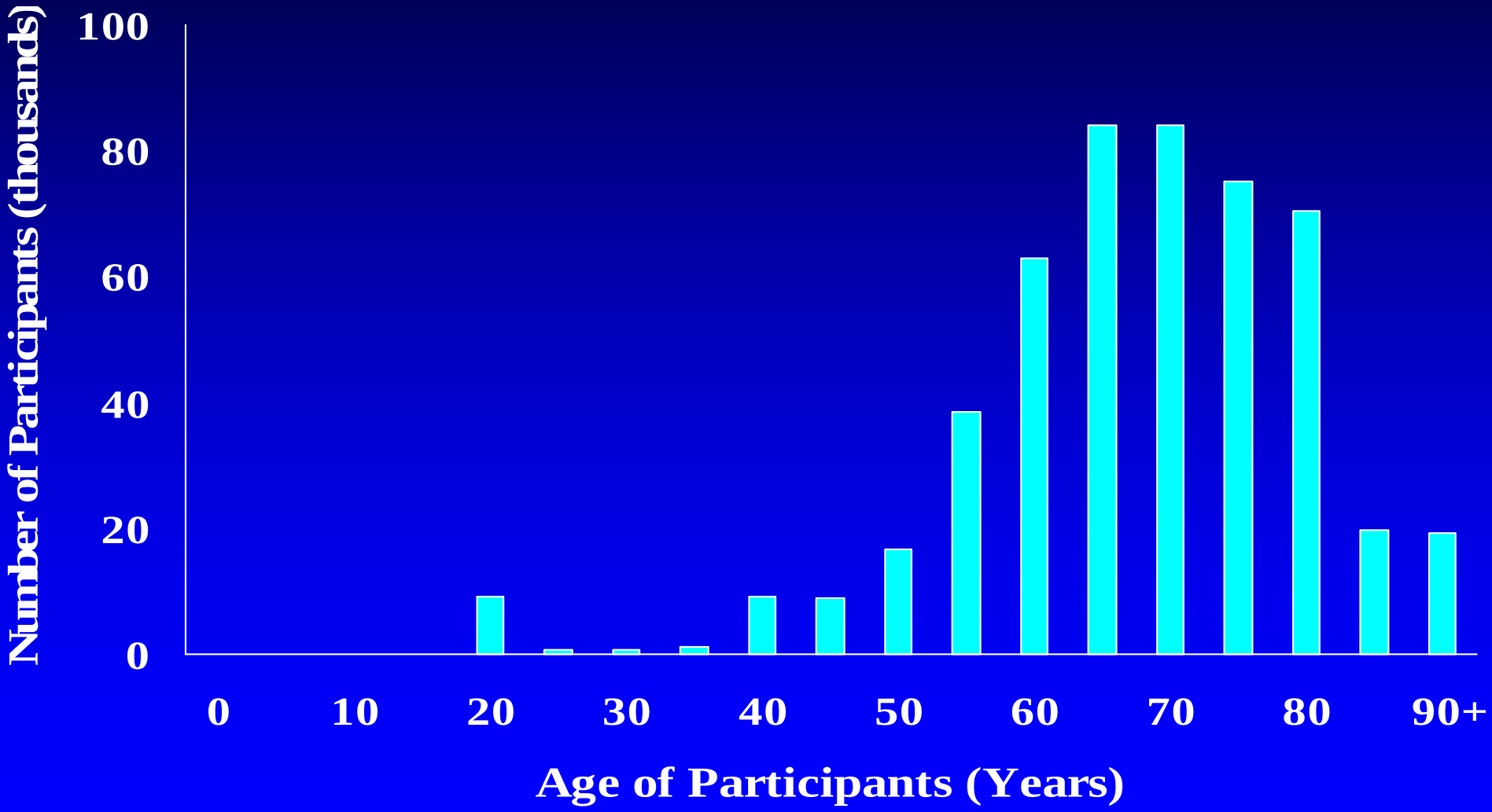
NOT-OD-04-041
May 5 – May 28, 2004

NOT-OD-04-046
June 1- June 30, 2004

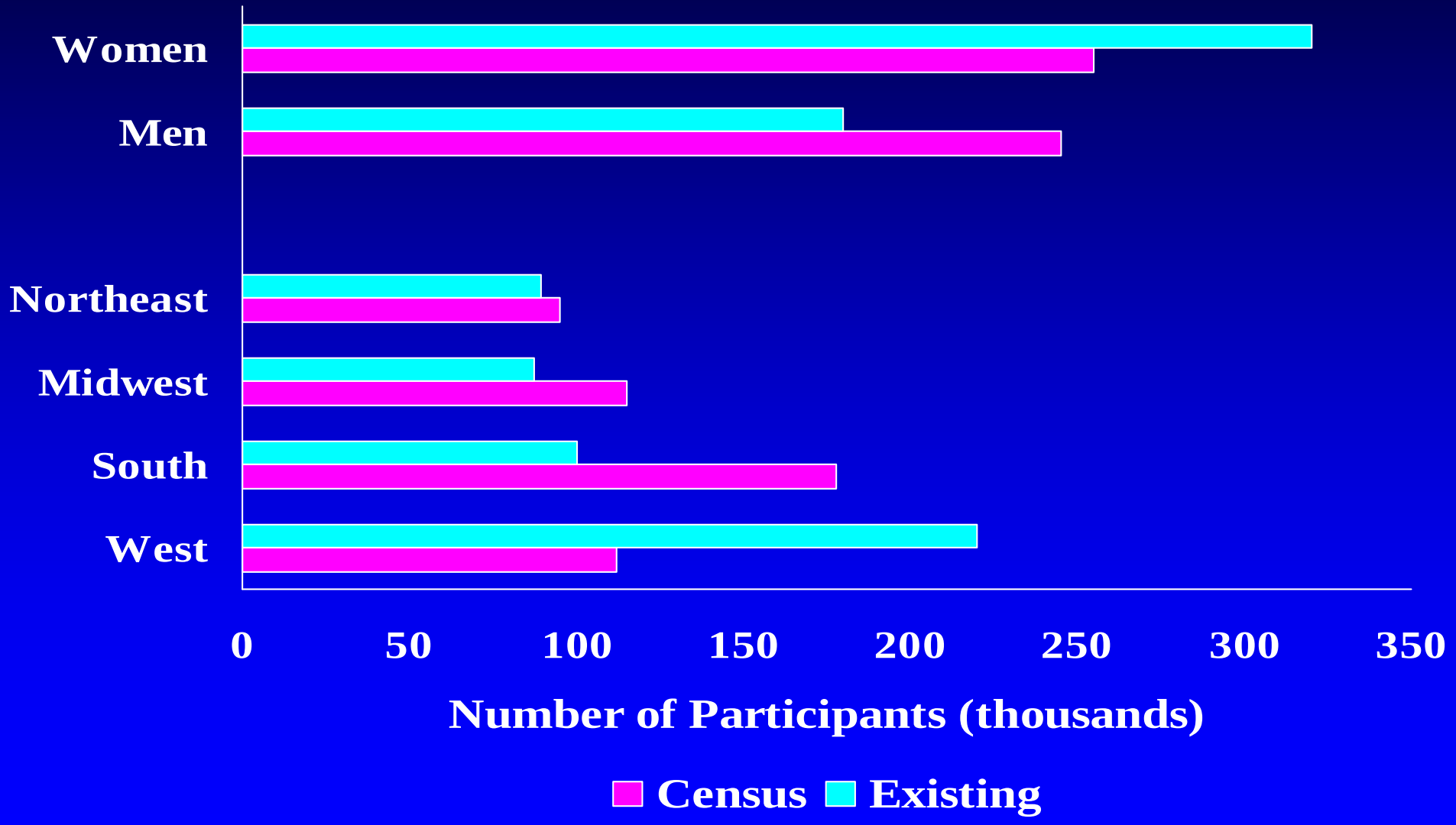
ESTIMATED AGE DISTRIBUTION OF REPRESENTATIVE US COHORT (2000 CENSUS)



ESTIMATED AGE DISTRIBUTION OF EXISTING NIH-FUNDED COHORTS



PROJECTED SEX AND REGIONAL DISTRIBUTION OF EXISTING COHORTS AND US CENSUS



PROJECTED EDUCATION DISTRIBUTION OF EXISTING COHORTS AND US CENSUS (Age ≥ 25)