U.S. Department of Health and Human Services

National Institutes of Health

National Heart, Lung, and Blood Institute

THE NIH INITIATIVES AND PLANS FOR BIOBANK STUDIES

NATIONAL HEART, LUNG, AND BLOOD INSTITUTE
NATIONAL HUMAN GENOME RESEARCH INSTITUTE

NATIONAL INSTITUTES OF HEALTH

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

Teri A. Manolio, M.D., Ph.D.
Director, Epidemiology and Biometry Program
Division of Epidemiology and Clinical Applications

OBJECTIVES

- Describe desirable characteristics of large US cohort study of genes and environment
- Outline possible designs for such a study
- Describe strengths and weaknesses of other study designs (existing cohorts, case-control studies)
- Outline some key priorities for large cohort studies and biobanks

insight commentary

The case for a US prospective cohort study of genes and environment

Francis S. Collins

National Human Genome Research Institute, National Institutes of Health, Building 31, Room 4B09, MSC 2152, 31 Center Drive, Bethesda, Maryland 20892-2152, USA (e-mail: fc23a@nih.gov)

Information from the Human Genome Project will be vital for defining the genetic and environmental factors that contribute to health and disease. Well-designed case–control studies of people with and without a particular disease are essential for this, but rigorous and unbiased conclusions about the causes of diseases and their population-wide impact will require a representative population to be monitored over time (a prospective cohort study). The time is right for the United States to consider such a project.

Identification of the genetic and environmental factors that contribute to health, disease and response to treatment is essential for the reduction of illness. This, of course, is the primary goal of biomedical research. Several auspicious recent developments suggest that progress in this area could be quite rapid. The sequence of the human genome[1,2] and increasing information about the genome's function have provided a robust foundation for the investigation of human health and disease. Likewise, results from the exploration of human genetic environmental exposure have improved. These techniques promise to extend the range of epidemiological investigation[5]. There is growing recognition that a change in the environment, in combination with genetic disposition, has produced most recent epidemics of chronic disease, and may hold the key for reversing the course of some diseases[6]. For example, consider the interaction of presumed famine-protective genetic predispositions with a modern environment in which there is a ready availability of excess calories. This has probably contributed to the current obesity epidemic

Collins FS, *Nature* 2004; 429:475-477.

NEED FOR LARGE US COHORT STUDY OF GENES AND ENVIRONMENT

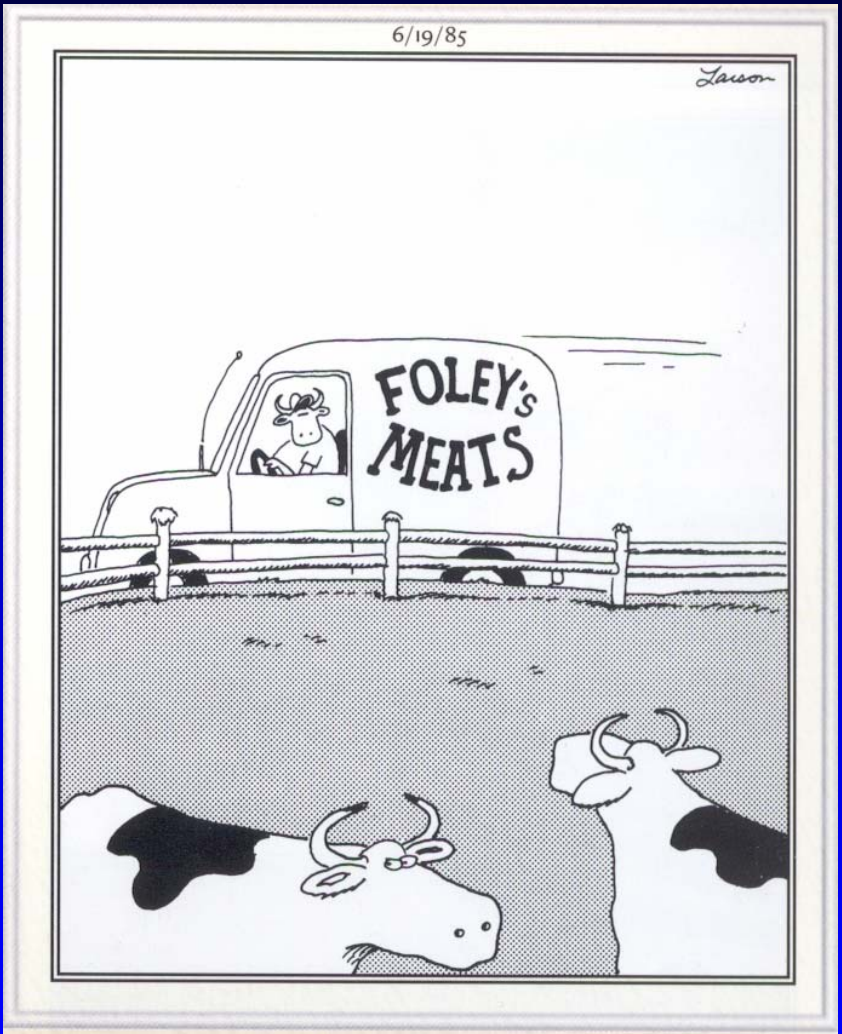
Identifying and reducing disease risk depends on unbiased determination of:

- quantitative contributions of environmental and genetic factors
- interactions among them
- complex interplay among disorders sharing common risk factors (such as heart disease, hypertension, and diabetes)

Replication of associations and estimation of their magnitude, consistency, and temporality best obtained through prospective, population-based cohort studies

Collins FS, *Nature* 2004; 429:475-477.

CONFLICT OF INTEREST DISCLOSURE



Eventually, Murray took the job—but his friends never did speak to him again.

Larson, G.
The Complete Far Side, 2003

DESIRABLE CHARACTERISTICS OF LARGE US COHORT STUDY

- Large sample size
- Full representation of minority groups
- Broad range of ages
- Broad range of genetic backgrounds and environmental exposures
- Family-based recruitment for at least part of the cohort to control for population stratification
- Broad array of clinical and laboratory data, regular follow up for events, additional exposure assessment

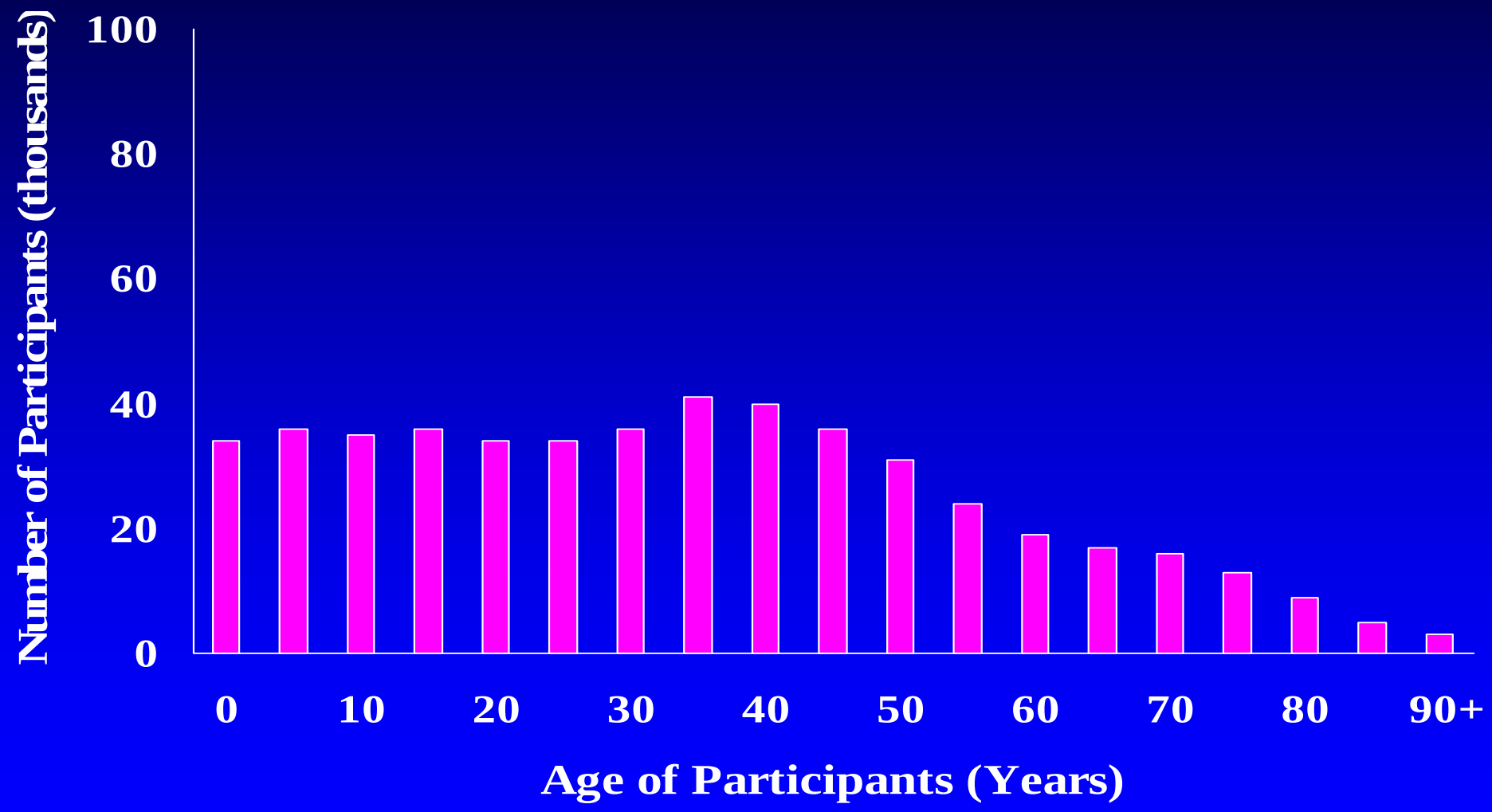
After Collins FS, *Nature* 2004; 429:475-477.

DESIRABLE CHARACTERISTICS OF LARGE US COHORT STUDY (continued)

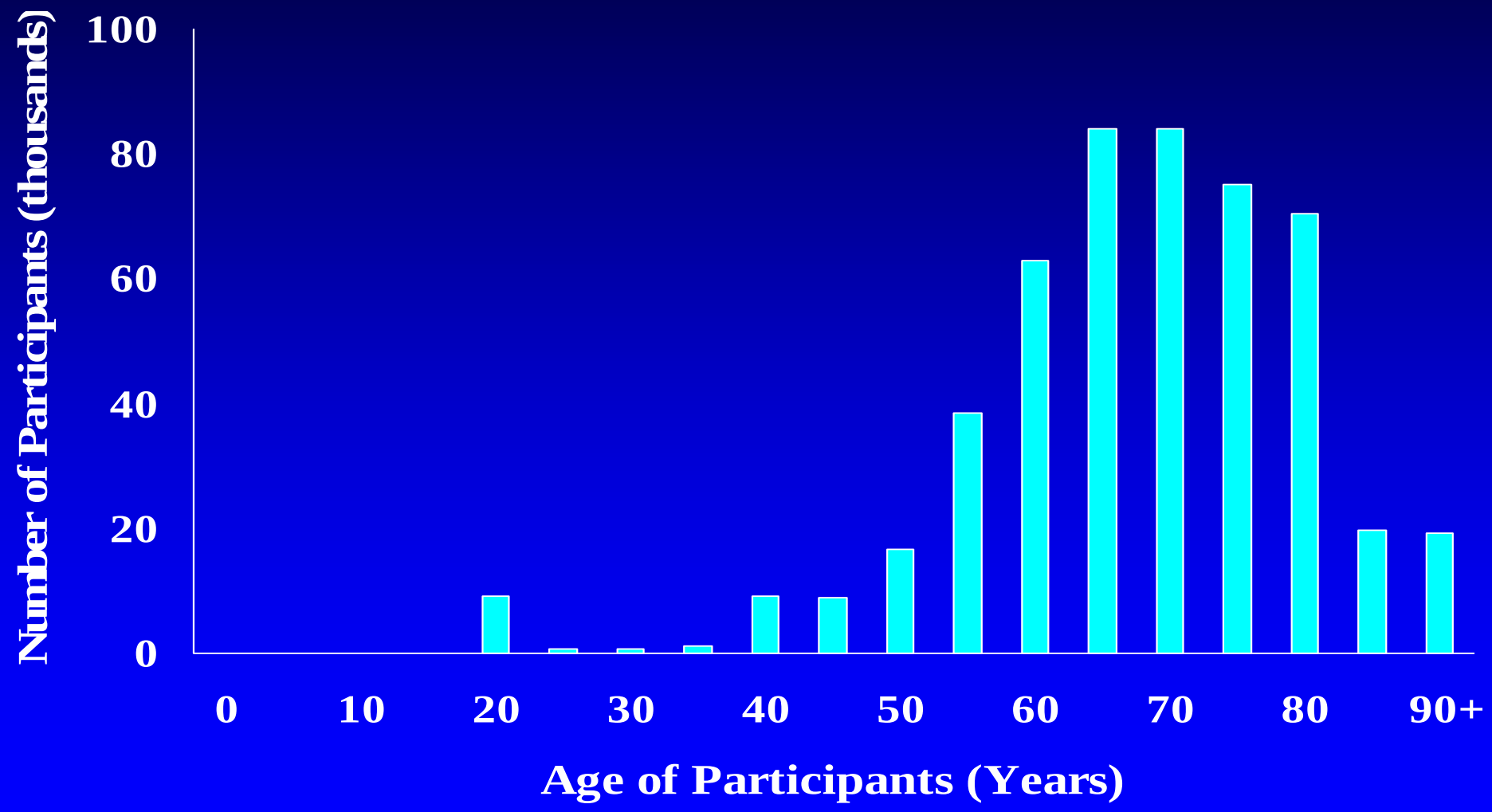
- Technologically advanced dietary, lifestyle, and environmental exposure data
- Collection and storage of biological specimens
- Sophisticated data management system
- Access to materials and data by all researchers
- Goals should not be “hypothesis-limited”
- Comprehensive community engagement from the outset
- State of the art (?dynamic) consent to allow multiple uses of data and regular feedback to participants

After Collins FS, *Nature* 2004; 429:475-477.

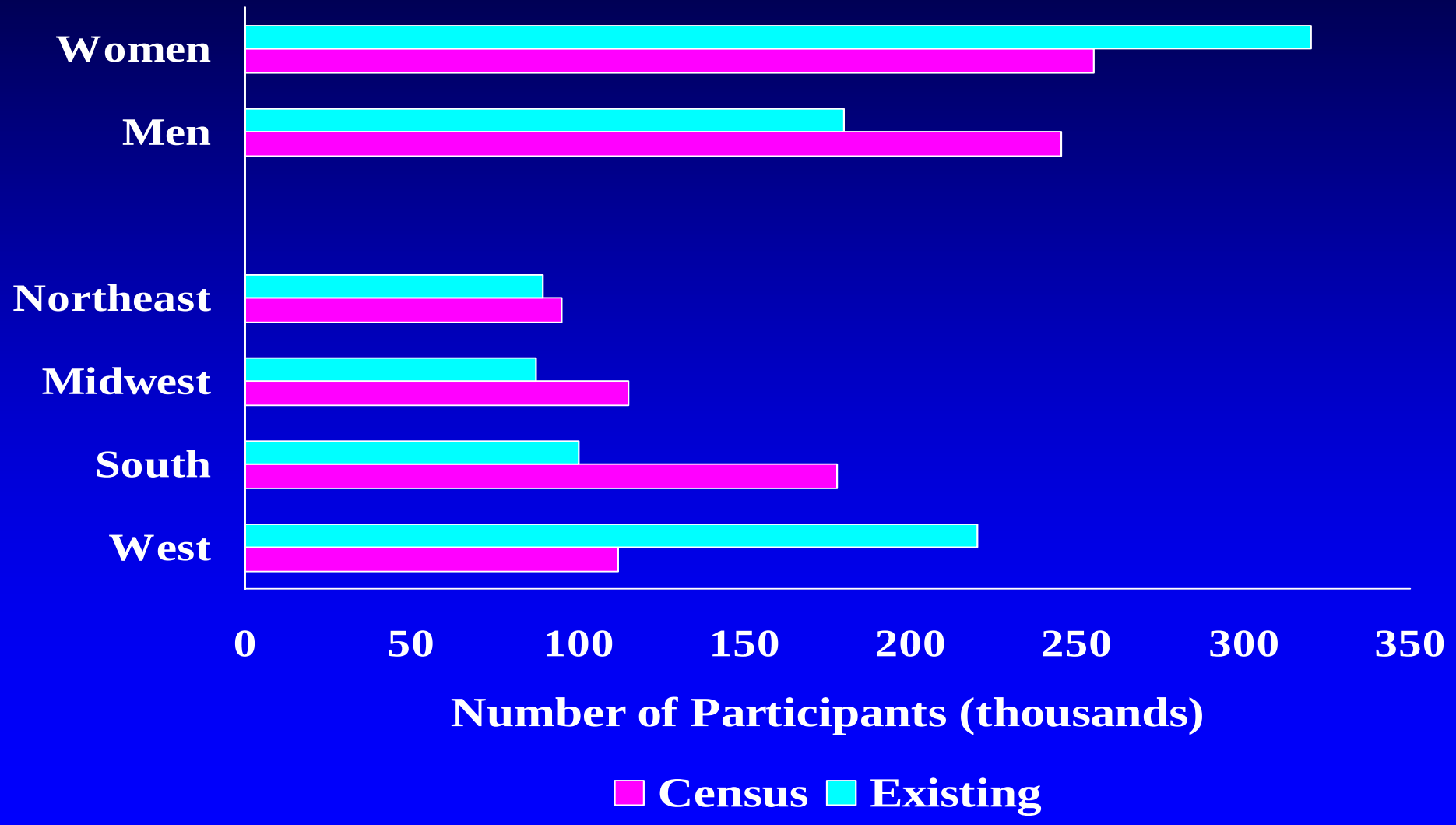
ESTIMATED AGE DISTRIBUTION OF REPRESENTATIVE US COHORT (2000 CENSUS)



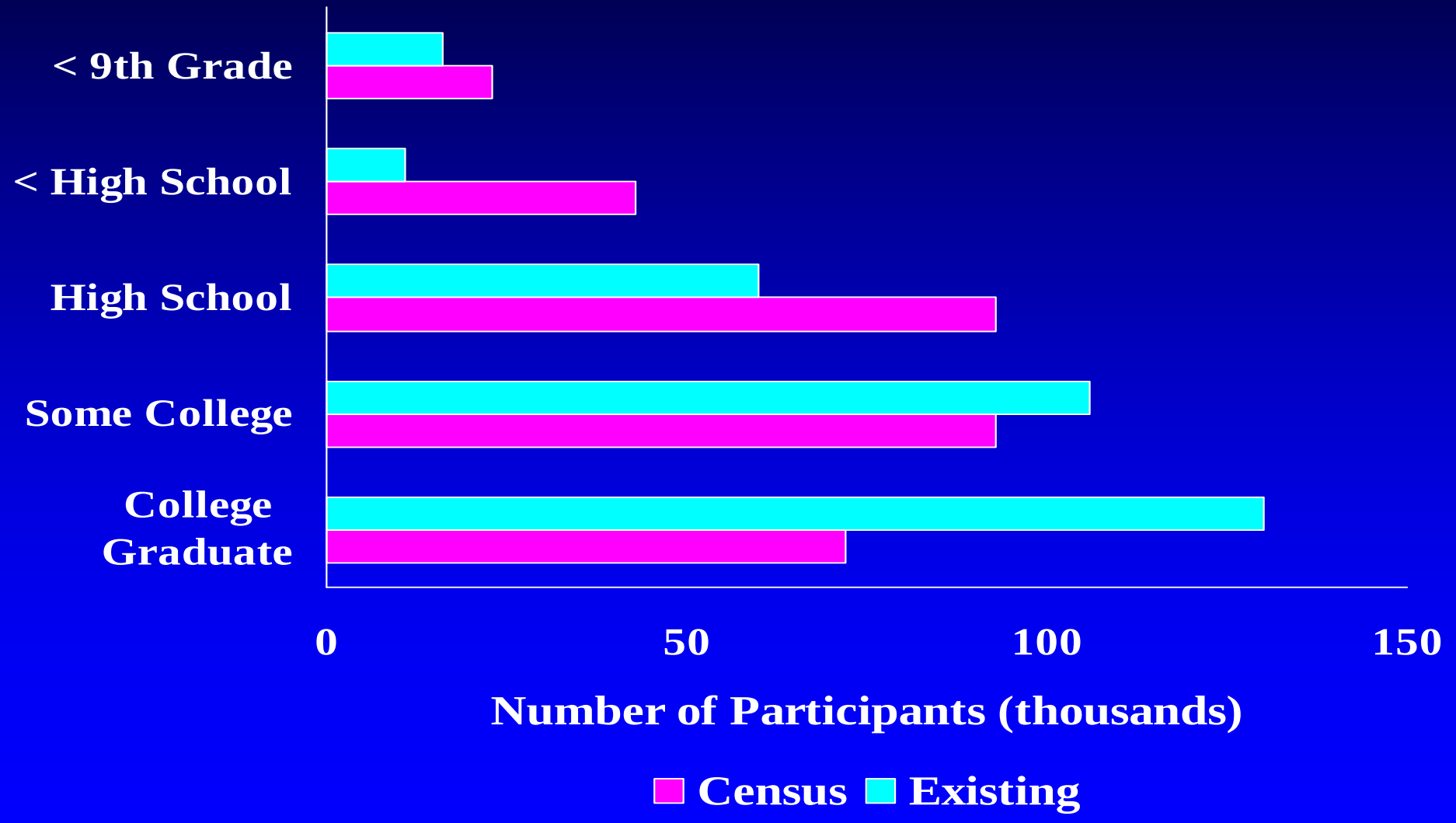
ESTIMATED AGE DISTRIBUTION OF EXISTING NIH-FUNDED COHORTS



PROJECTED SEX AND REGIONAL DISTRIBUTION OF EXISTING COHORTS AND US CENSUS



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



SAMPLE SIZE AND POWER ESTIMATES

- Primary goal to assess critical gene-by-gene and gene-by-environment interactions
- Minimum number of cases needed to detect desired relative risk estimated; assumed 2 matched controls per case
- Range of allele frequencies, environmental exposures, dominance models shown
- Population-based incidence estimates where available
- Number of new cases in cohort of size 200,000, 500,000, or 1 million
- Estimated minimum relative risk with 80% power and Type I error = 0.0001

POPULATION-BASED COHORT STUDIES

- Definition: prospective investigation of representative sample of population followed for development of specified endpoints
- Purpose: to identify risk factors predisposing to development of disease in the general population, particularly risk factors:

  affected by disease, treatment, lifestyle changes
  subject to imperfect or biased recall
  with hypothesized early pathogenic effect

- Complement other epidemiologic study designs:

  surveillance studies
  cross-sectional surveys
  case-control studies
  clinical epidemiology studies

PROS AND CONS OF COHORT STUDIES

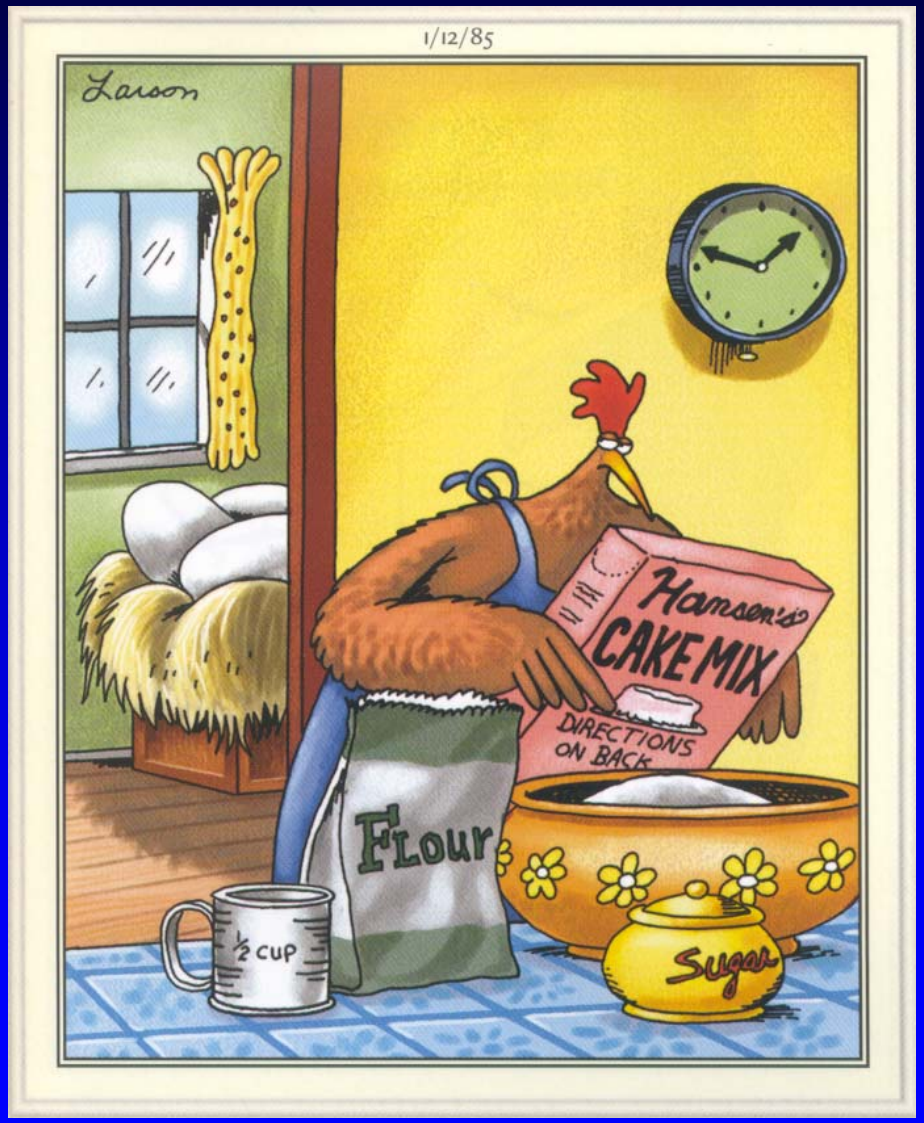
DISADVANTAGES

- They are expensive.
- They take a long time.
- They are very broad-based.

ADVANTAGES

- They provide risk information obtainable through no other means.
- They are understandable to the public and media.
- They identify modifiable risk factors for potential preventive interventions.

CONFLICT OF INTEREST DISCLOSURE



Larson, G.
The Complete Far Side, 2003

MAJOR NHLBI COHORT STUDIES

Study	N	Age	Entry	Minorities
Framingham Cohort	5,209	28-62	1948-50	--
Framingham Offspring	5,124	20-74	1971-75	--
Framingham Gen3	~ 4,000	20-60	2002-04	--
Honolulu Heart	8,006	46-68	1965-68	100% JA
CARDIA	5,115	18-30	1985-86	52% AA
ARIC	15,787	45-64	1985-87	27% AA
CHS	5,888	65-100	1989-90	16% AA
Strong Heart	4,549	45-74	1989-91	100% AI
WHI	161,809	50-79	1993-98	18% multiple
MESA	6,749	45-84	2000-02	28% AA, 22% HA, 12% CA
Jackson Heart	5,308	35-84	2000-04	100% AA
Hispanic Cohort	16,000	35-84	2006-10	100% HA

BASIC ASSUMPTIONS FOR BIAS-FREE CASE-CONTROL STUDY

- Cases are representative of all persons who develop the disease/condition
- Controls are representative of the general “healthy” population who do not develop the disease
- Collection of risk factor and exposure information is the same for cases and controls

PROS AND CONS OF CASE-CONTROL STUDIES

ADVANTAGES

- May be the only way to study rare diseases or those of long latency
- Existing records can occasionally be used if risk factor data collected independent of disease status
- Can study multiple etiologic factors simultaneously
- May be less time-consuming and expensive
- If assumptions met, inferences are reliable

PROS AND CONS OF CASE-CONTROL STUDIES

DISADVANTAGES

- Relies on recall or records for information on past exposures; validation can be difficult or impossible
- Selection of appropriate comparison group may be difficult
- Multiple biases may give spurious evidence of association between risk factor and disease
- Usually cannot study rare exposures
- Temporal relationship between exposure and disease can be difficult to determine

"BUT," THEY SAY, "*THIS* IS GENETICS!"

(you dumb epidemiologist)

"*THIS* IS DIFFERENT!"

- Genes are measured the same way in cases and controls
- Information on key exposure is easy to validate
- No recall or reporting involved
- Temporal relationship between genes and disease is piece of cake

"BUT," I SAY,

- Bias-free ascertainment of cases and controls is still major concern; cases in most clinical series unlikely to be representative
- Assessment of risk modifiers or gene-environment interactions is likely to be incomplete or flawed

CASE-CONTROL STUDIES AND RARE DISEASES

- For a disease with incidence of 8 cases per 1,000 among unexposed, cohort study would require 3,889 exposed and 3,889 unexposed persons to detect two-fold increase in risk
- Case-control study would require 188 cases and 188 controls, assuming 30% exposure
- For disease with incidence of 2 cases per 1,000 among unexposed, would need 15,700 exposed and 15,700 unexposed to detect two-fold risk
- Case-control study would *still require only 188 cases and 188 controls*

Schlesselman JJ. *Case-Control Studies*, 1982.

SO WHAT'S A MOTHER TO DO?

- "Nesting" a case-control study within a prospective cohort may provide the best of both worlds
- Large proportion of cohort members who do not develop disease provide little incremental information
- If exposure information can be collected and stored for later measurement, can wait for cases to accrue and then measure exposures in limited sample of non-cases
  - stored biologic samples
  - stored images
- Can be expanded to "case-cohort" concept with representative sample of cohort, regardless of disease status, used for multiple comparisons

LARGE COHORT STUDIES OF GENES AND ENVIRONMENT: PRIORITIES

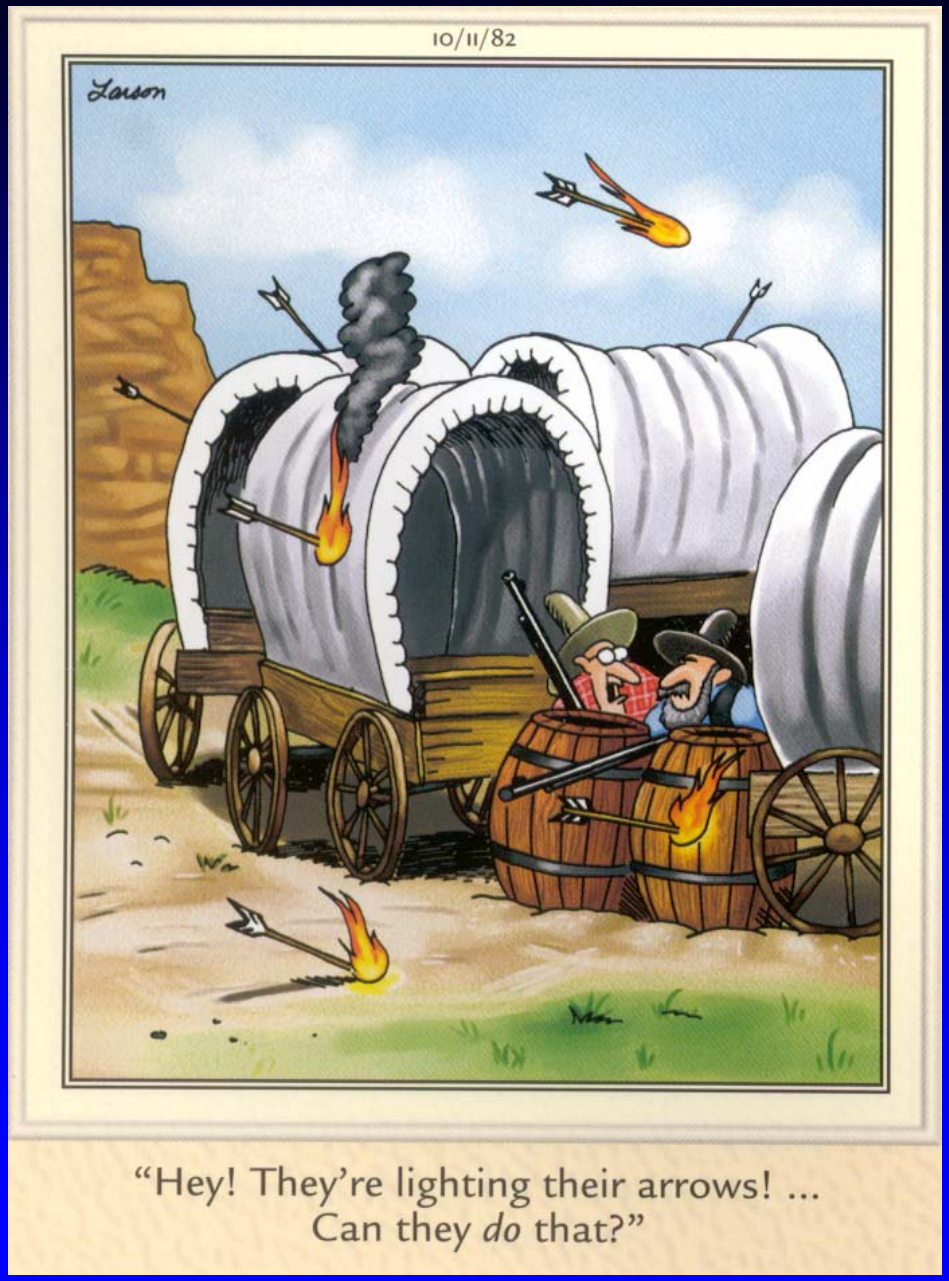
- Promote sharing of protocols and data

LARGE COHORT STUDIES OF GENES AND ENVIRONMENT: PRIORITIES

- Promote sharing of protocols and data
- Promote sharing of protocols and data

LARGE COHORT STUDIES OF GENES AND ENVIRONMENT: PRIORITIES

- Promote sharing of protocols and data
- Promote sharing of protocols and data
- Promote sharing of protocols and data



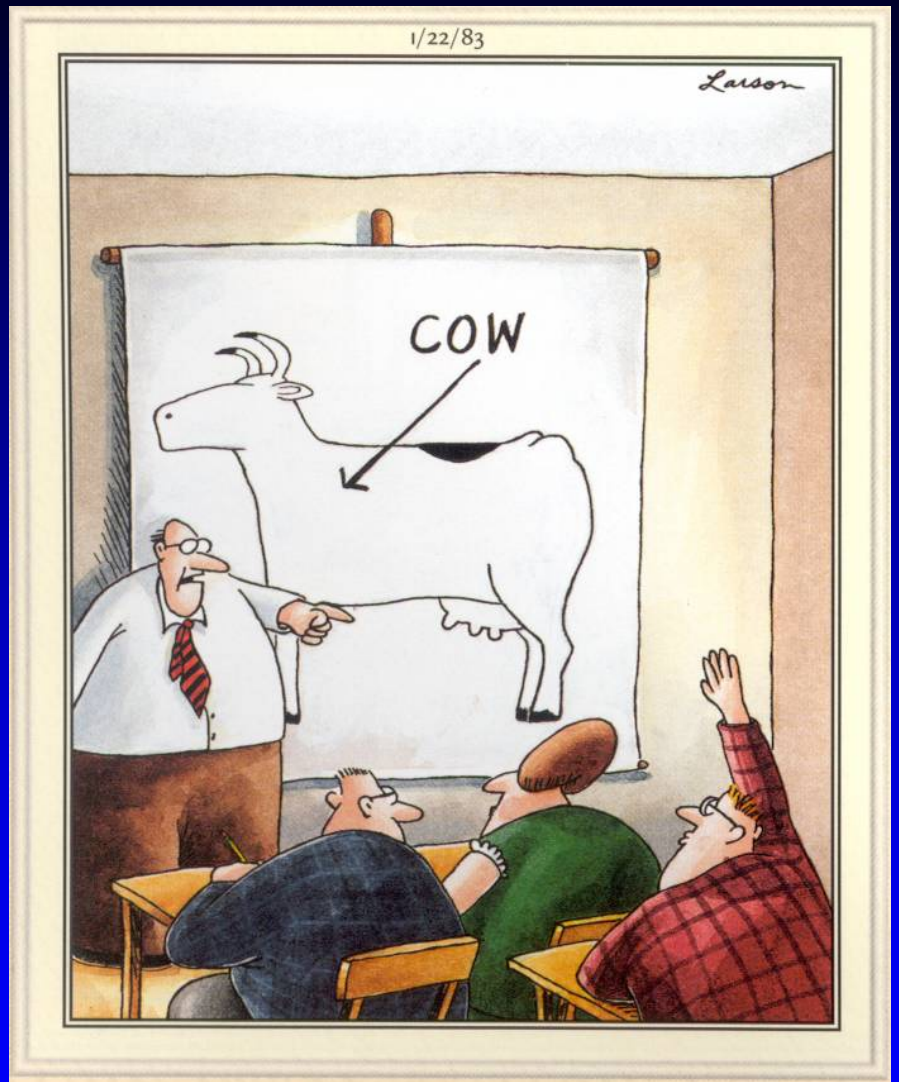
Larson, G.
The Complete Far Side, 2003

LARGE COHORT STUDIES OF GENES AND ENVIRONMENT: PRIORITIES

- Promote sharing of protocols and data
- Ensure core of phenotypic and exposure information collected in standardized way in exchangeable formats

- medical history	- diet, physical activity	- lab
- medication use	- anthopometry	- BP
- school/work absence	- physical performance	- PFT
- self-reported health	- cognitive function	- outcomes!
- occupational history	- residence history (GIS)	

- Genotype and correlate core set of known variants and "anonymous markers" across studies



"Yes ... I believe there's a question there in the back."

Larson, G.
The Complete Far Side, 2003

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

MAJOR QUESTIONS AND RESPONSES (150 RESPONDENTS)

1. New cohort vs. existing cohorts
2. Desirable characteristics of large US cohort study
3. Family structures recommended for inclusion
4. Issues relevant to power
5. Other comments

<u>If responsible for existing study</u>:

6. Likelihood of participation and contribution of data
7. Likelihood of making data available outside this effort

ADVANTAGES OF NEW COHORT

- Design: based on needs of study rather than convenience; get it right from the start
- State of art: use up-to-date technology, address current health concerns
- Consistent protocol: avoid lowest common denominator
- Poolability/survivorship: easier to pool on genetics than environment?
- Consent: more straightforward, up-to-date, avoiding complexity of many changes over time

ADVANTAGES OF NEW COHORT (2)

- Multiple outcomes: built in from start
- Free and open access: establish up front; consider separating functions that store and distribute from those that collect and analyze
- Biologic specimens: fresh, high-quality, suitable for proteomics or RNA analysis
- Diversity
- Younger ages: most existing cohorts middle age or older

ADVANTAGES OF EXISTING COHORTS

- Saves time/money: usefully supplement in cost-effective way, leverage existing investment
- Experience and expertise: already shown can collect high quality data
- Recruitment: may have higher response rate
- Community responsiveness: relationships with communities already established
- IRB and institution-specific requirements: time-consuming, iterative process, already worked out
- Valuable ongoing work: don’t be too quick to abandon

DESIRABLE CHARACTERISTICS OF LARGE COHORT STUDY

- Representative, rigorously population-based
- Diverse regarding:

  - age (a few preferred younger and a few, older, cohort)
  - race/ethnicity (only 1 urged ethnically homogeneous)
  - sex
  - SES
  - region, urban/rural
  - occupation
  - sexual orientation
  - multiple diseases
  - dietary/other environmental exposures (vs. 1 homogenous)

DESIRABLE CHARACTERISTICS OF LARGE COHORT STUDY

- Uniform, high quality phenotypic characterization
- High quality lifestyle, diet/activity, occupational, environmental exposure data
- Completeness of follow-up, with documentation
- Flexible but robust infrastructure to accommodate variety of analyses
- Close involvement of community members in design, execution, communication

LARGE-SCALE GENOTYPING OF NHLBI COHORTS

- Standardized assessment of phenotypes and exposures is primary emphasis of population-based cohort studies
- NHLBI cohort studies have typed numerous candidate genes, but gene selection largely driven by investigators' interests, methods for genotyping varied, few variants typed at a time
- Wealth of phenotypic information defies any group of investigators to exploit fully; efforts to ensure open access and promote data sharing have had limited success

LARGE-SCALE GENOTYPING OF NHLBI COHORTS

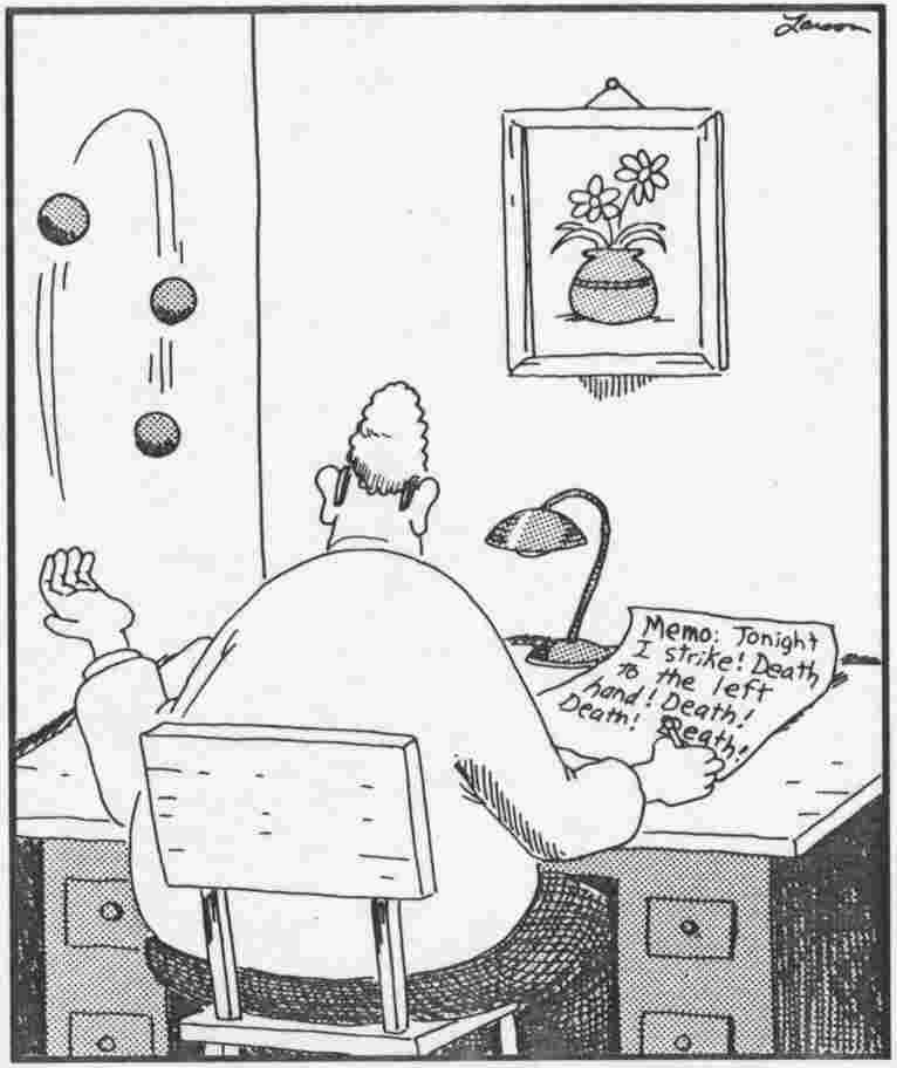
- Consider genotyping ~10 SNPs in ~1,500 candidate genes in ~50,000 cohort study participants
- Make data rapidly and widely available to IRB-approved investigators completing confidentiality agreement
- Consider genome-wide association study of ~300,000 SNPs in 500 cases and 1,000 controls, providing 80% power to detect allele of 20% frequency carrying relative risk of 1.7 with type I error < 0.0001
- Challenges in combining existing phenotypic data will be substantial; consider development of NHLBI-wide common database similar to caBIG

NATIONAL CANCER INSTITUTE

caBIG cancer Biomedical Informatics Grid

- informatics infrastructure to connect research teams, enable better development and sharing of tools and data in open environment with common standards
- tools include standards-based, components-based clinical trial management systems
- standards and tools will support common vocabularies, data elements, and a unifying architecture

http://cabig.nci.nih.gov/overview/



Innocent and carefree, Stuart's left hand didn't know what the right was doing.

THE FAR SIDE

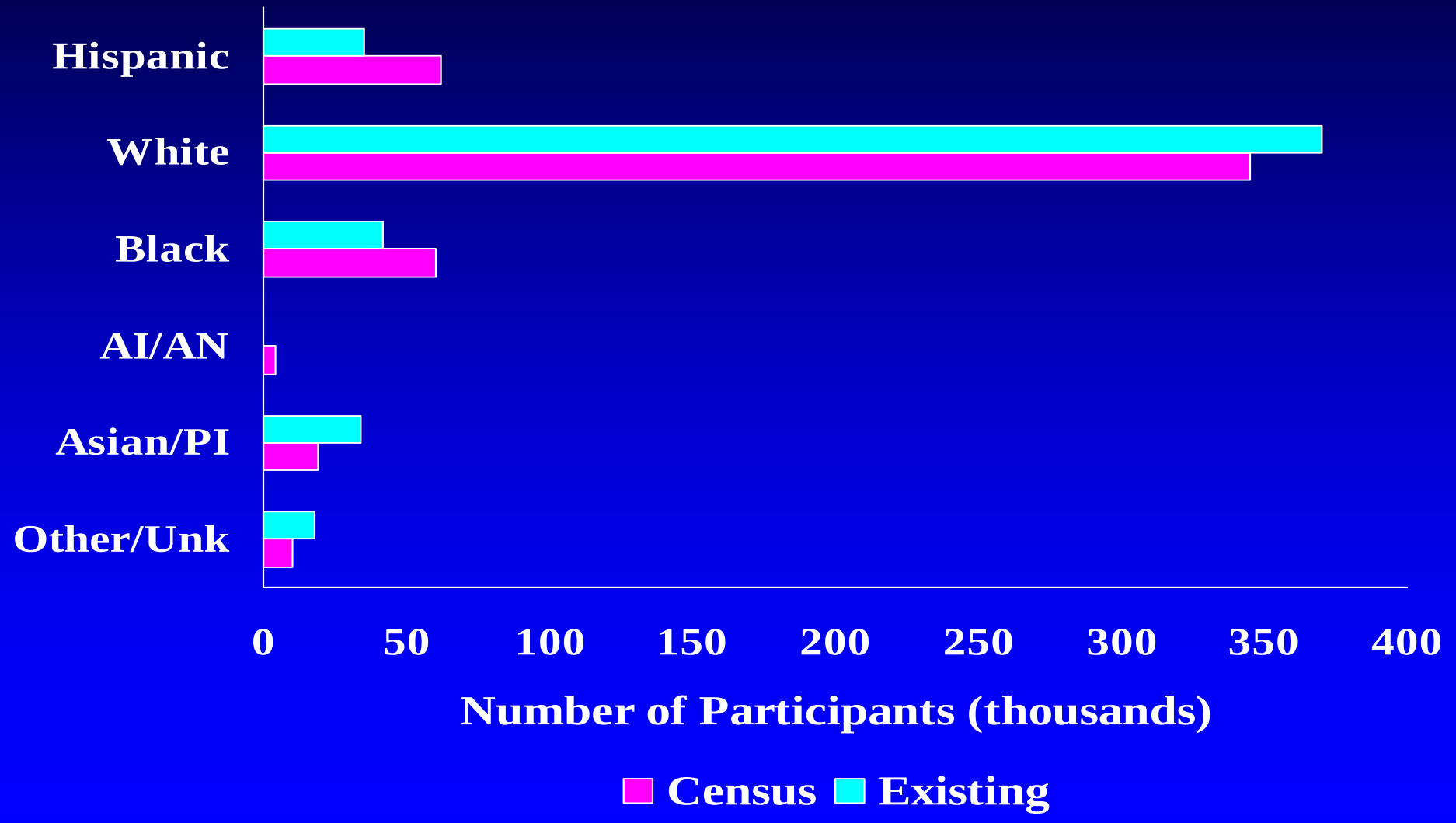
JULY

29

THURSDAY

NEED FOR COMMON DATA BASE IN NHLBI COHORT STUDIES

- promote consistent data collection
- eliminate unneeded or redundant data collection
- reduce or eliminate need for each new study to develop its own data collection system
- promote consistent reporting and analysis across studies
- reduce the possibility of error related to data translation and transmission
- facilitate data sharing

NIH Roadmap: Re-engineering the Clinical Research Enterprise, Clinical Research Networks; http://nihroadmap.nih.gov/

PROJECTED RACE/ETHNIC DISTRIBUTION OF EXISTING COHORTS AND US CENSUS