Federal Perspectives on the Need for a Large Population Study
Alan Guttmacher, M.D.

DR. TUCKSON: Let me invite Alan Guttmacher from the National Human Genome Research Institute, who has been very active in trying to get something launched themselves.

DR. GUTTMACHER: It's a real pleasure to be here and talk with the committee about something that I think that many of you have expressed interest about. The committee has heard something over the last six to nine months about a group that was meeting at the NIH to look into the really scientific questions about a possible large U.S.-based gene/environment.

Actually, I'm going to quibble with my own title slide. Even though we call it study because AGES is an easy acronym to be able to refer to, this as a sort of working concept, it is really more of a resource than a study. I think for study, the word "study" to many people implies a kind of controlled thing that is really hypothesis-driven. You have a specific hypothesis, and you're going to do a study to answer that hypothesis. We think of this as more hypothesis informed rather than driven. That is, it should be a large resource available to, as you'll see in a moment, basically the entire research community to be able to answer a series of very interesting hypotheses and questions.

You have to have some sort of exemplar or hypotheses as you design something like this, because you might want to say gee, if it couldn't handle the following kind of question, why bother having this resource? But on the other hand, if we're thinking about large, longitudinal studies, one of the things we kept in our minds as we thought about this was they obviously will be providing data for years to come.

If, for instance, using the model, as many do when they think about these sorts of studies at Framingham, if you had gone back to the original days of the Framingham study and asked them to define the hypotheses which they would be using the Framingham study to answer in the year 2005, we would have done a pretty poor job of that.

We think the same kind of thing for these large longitudinal studies. You have heard this from many of the speakers before. The one needs to really be thinking very far forward, and therefore really thinking beyond our ability to think and to be aware of that as we go into it.

So obviously there are various kinds of approaches to discovering and quantitating the genetic and environmental contributions of disease risk. We have been talking about those all morning. Case-control studies and prospective population-based cohort studies. Case-controlled studies are great, and that's perhaps the most important part of this slide, that even those of us thinking about this are clearly cognizant of the idea that case-controlled studies are wonderful things, and that we need to continue to have those for biomedical research. But there are some things they can't do.

Teri Manolio and others talked about some of the things that they could do and could not do. Amongst the aspects that Teri talked about, or particularly emphasized, are the bias towards the more severe end of the disease spectrum. This recall bias which Teri spoke about was in terms of both environmental exposures and family history.

For instance, there are several here who have done some teratology research over the years. We certainly all have learned the lesson that cases tend to have different memories from controls. Very importantly, the inability, using case-control studies, the limited ability to identify

predictive biomarkers that signal the future onset of disease and to have good information about those controls before they become cases, because of course we want to have those early biomarkers.

Now, as you well know, we've heard about many of the other countries that are planning large population-based studies of genes, environments, and health. Why doesn't that suffice? Those are going to be wonderful studies. But there are some problems for those of us in the U.S. in terms of utilizing these.

These include, and there are others besides these three, but perhaps the three major ones that other countries do not reflect are the population groups, no matter how one defines population groups. But the population groups in the U.S., particularly those very groups that seem to be at present most involved with having health disparities.

Other countries do not reflect the environmental factors found in the U.S. This will vary from country to country in how well that reflection is found, but it is not a full reflection of some of the environmental factors in the U.S. Be they the physical environment, social environment, or other kinds of environment.

Also this question about access of particularly U.S. researchers, but researchers in general, to data from other country studies will, as you've heard, be limited. So for all of those reasons, we thought there was reason to think about a U.S.-based study. Many of you will know about this, it is available in the materials. I think it is in everyone's binder that Frances wrote an article last summer, the case for U.S. prospective Cohort Study in Genes and Environment, which I would refer you to it. It outlines many of the reasons for thinking about this.

A working group was convened, and these are the members of the core working group. I should also add that Teri Manolio's name does not appear in this. That's because she, along with Frances and I, were surfing the NIH perspective helping to sort of pull this together and organize it. Teri was a very active participant. She mentioned before being honest about her relationship with the Iceland group. I'm not sure why she refused to mention her relationship with our group. Perhaps she was a little worried about what I might say.

(Laughter.)

DR. GUTTMACHER: It shows how well she knows our group. Besides these folks, there were a number of subgroups, which you'll see here, which included another 50 people. So there were a total of about 60 folks from both the United States and from outside the United States involved in helping us think this out over the last, as I said, six to nine months.

So what are the major recommendations? I would emphasize major. The more detailed kind of information, I'll tell you at the end of the talk how to find that. But let me just sort of skate through some of the major ones since time is limited.

At the end of the day, the feeling was that cohorts should be chosen to match the most recent U.S. Census on six different characteristics. In terms of age, in terms of sex, in terms of race/ethnicity, in terms of geographic region, in terms of education, and in terms of urban versus rural residence.

It was also felt that the household should be the primary sampling unit, and that roughly 30 percent of cases should consist of biologically related individuals. I would like to point out that's

not a floor, it's a target. In fact, there is an advantage to holding it not much above that, as well as an advantage to getting somewhere towards that.

It was also felt that the cohort should be a significant size to achieve adequate power for most common diseases and quantitative traits. If that does not seem obvious to you by now, you haven't paid very much attention this morning.

What does significant mean? Well, we did a number of various kinds of models to look at it. This is one that looks at the minimal detectable odds ratio contributed by a genetic variant after five years of follow-up, looking at various diseases in terms of their incidence per 100,000 in the population per year, with the assumptions up there of 80 percent power, and looked at various cohort sizes, 200,000, 500,000, and 1,000,000. To no one's great surprise, the larger the cohort, the more data you get.

We also looked at of course because we weren't just interested in this alone, but also looked at minimum detectable environmental odds ratio after five years of follow-up for the same spectrum of disorders in terms of incidence.

Finally, we looked at it in terms of gene by environment interaction, which of course is perhaps what we'd be most interested in after a five-year follow-up. Now, there are a number of assumptions. Part of what this really presents is that there is no sudden sweet spot or something. There is no number where you suddenly say gee, this is a number you should get. Obviously the smaller the study, the easier to do. So if there is some magic number beyond which you don't get much added information if you get larger, no, so any kind of type of design of this is going to weigh the scientific possibilities versus some of the budgetary constraints.

What else did the group think about? Well, clinical exam obviously would be important. We thought that a baseline assessment should be done, which should be limited to four hours for various logistic reasons, that a core group of variables should be collected on all participants, and other variables that would be
age-specific to the participants.

Again, remember, the age of this resource would reflect the ages that we see in the U.S. population, that biological specimens should be collected, laboratory measurements done upon them, the specimens should be stored, the genotype and DNA sequencing would be done.

In terms of follow-up, that there would be telephone or email contact every six months, and that reexamination should be carried out every four-year periodicity.

Public consultation. We should also add that in here. Not just extensive, but early and extensive. There was a feeling that for something like this to work, for lots of reasons, there has to be, as many people alluded to before, that participants are truly participants, that they feel and deserve to feel a sense of ownership of this, that this would include various kinds of town meetings and focus groups before one even got started.

There should be an open-ended, informed consent with an encrypted database to protect privacy and confidentiality to the degree that one can protect it, but obviously being completely honest with participants about the limits of any protections. A central IRB would be highly advantageous, which is obviously something that many would aspire to. It would not be unchallenging to pull off.

Data should be immediately accessible to all investigators who have IRB approval. I would like to underline this. This is perhaps a distinctive feature of this design. It is not unique, but certainly a very important part of this to us. That would not be something where a closed group of investigators would have access to the information, that much of what we were thinking about sort of came from a Human Genome Project-type model, and part of the power of the Human Genome Project was having data immediately accessible to as many investigators as possible.

Here one needs obviously to weigh that against various kinds of concerns for privacy and confidentiality of participants. We think by using IRB for approval, that one could pull that off.

So why do this now? Well, the urgency of discovery and validating these kinds of things, the same things that John and others have spoken about before. The opportunities to understand and address causes of health disparities, and also that we think this will be a powerful stimulus for technology development, as many of these kinds of population studies could be, we would like to use this to help do some of the work that Gil mentioned before about really driving innovation in terms of measurement of both environmental factors, as well as better describing phenotype with new technologies.

Also, the potential to reduce skyrocketing health care costs by understanding better the etiology of disease and people's response to treatment for disease.

Finally, I will mention to you that by the close of business today, I believe there will be a full report of that working group. We've been working hard to try to pull it together for this meeting. We believe by the end of the working day today, and since we are federal folks, the close of business means midnight. Sometime today. If you go to genome.gov, that is the website. if you go to genome.gov/13014436, you will see a full report of the working group.

DR. TUCKSON: Thank you, Alan. What we can probably do, and maybe with the support of our staff, we can just get a little handout of that so that people will have that available. Thank you very much.