Federal Perspectives on the Need for a Large Population Study
Muin J. Khoury, M.D., Ph.D.

DR. TUCKSON: Muin Khoury, if you would give us the perspective from the Centers for Disease Control and Prevention. Then we will move expeditiously to the panel discussion that will be led by Hunt.

DR. KHOURY: Good morning. I guess I'm Speaker Number 10 this morning. By this time, you're all hungry and tired, and you've heard it all. So I'll try to be very quick so that we can have some discussion.

I'll try to offer you a bit of a global perspective on how we can go about collaborating, whether it is case-controlled cohort studies, or what have you. A lot of what I have to say is in this letter of correspondence to Nature Genetics last year. But because of the format, I had to condense it to about 600 words. But a full report of this is available on our website.

Now, I have three messages to you this morning. They will reflect partly my own philosophy in what CDC is doing with global collaboration with many of the people you've heard from before, and I mentioned specifically a couple of things.

The three messages this morning is that global collaboration in Biobank and population-based cohort studies is needed. We are beginning to see the elements of that with P3G, U.K. Biobank, and others. I firmly believe one cohort study in one country is not enough, no matter how big that study is, whether it has 1 million people or 2 million people.

You have seen some calculations from Alan Guttmacher earlier. They were based on measuring one gene and one exposure or gene/environment interaction. You could see those minimal detectable odds ratios creeping up as you begin to look at interactions. But if you are beginning to look at five or ten genes interacting with five or ten exposures, it is going to be quite challenging.

The second message I want to say this morning is that we need the process that integrates all of the human genome epidemiologic information, whether it comes from cohort studies, case-controlled studies, or other forms of studies. For the most part, most such data still come from case-control studies, and will for the foreseeable future. So we need to integrate that data as well.

Then the third, which I won't talk about today, is the need to link epidemiology with the evidence-based processes that use epidemiologic information for policy and practice. So there is a method to this madness. There is an epidemiologic approach that many of us have learned that applies not only to exposure, but genes. Because it is a huge problem literally, I decided to call it human genome epidemiology. Not because I have delusions of hugeness or anything, but because the problem is really huge on a practical scale.

What we deal with primarily these days is the processes of gene discovery, like the first speaker this morning who warned us that we need to kind of put on a different hat when we're talking about multifactorial diseases. We are not really discovering genes for diseases X, Y, and Z, but looking at how genetic variation, whether it is 10 million SNPs or just three SNPs or whatever, affect the risk of diseases.

Why do we need epidemiology? We need epidemiology to characterize what we have in the population, the prevalence of the gene variance, how they affect the burden of disease in terms of relative risks, absolute risks, and also the burden of disease. Then also characterize gene/gene and gene/environment interaction.

You have heard about all of these by now, and you are sick and tired of the different study designs. They all have their advantages and limitations. But there are also hybrid study designs. You can conduct a cohort study for which you can measure exposures retrospectively.

For example, if you had collected information from a newborn blood spot and have stored it for many years, you can go back to that blood spot and measure both genes and environment. So you can still do a case-controlled study having the antecedence of exposures measured before the case and controls were collected.

There are a couple of myths and stigmas about association studies that are in the literature. The term "association study" almost is like a dirty word in genetics. I think it is a function of the poor quality of association studies. Not because the field or the epidemiologic approach to association studies is bad. It is because the studies that are being done are really bad studies where the cases and controls come from different populations, and they are not even comparable, where you have both selection bias and all sorts of things.

Incidentally, both cohort studies and case-controlled studies are association studies. So there is that stigma that associates with that.

One thing I wanted to say here. Because of the lack of randomization, people talk about observation study as a second place class science. We don't determine who gets what allele. We are essentially randomized at miosis, or at birth. There is a movement, especially in Europe and the U.K., called the Mendelian randomization movement where it really takes the term "association study" and puts a randomized controlled clinical trial on it.

So basically it is randomizing people into Allele A and Allele B, and then look at the outcomes later. You don't choose which allele you get. It is just like you don't choose which drug you get from a controlled clinical trial. So we are taking the realm of association and making it closer to experimental design. We don't have time to talk about this.

Now, there is also this belief that cohort studies are inherently superior to case-controlled studies. Or case-controlled studies are inherently inferior to cohort studies. I am here to tell you that a well designed population-based case-controlled study is far more superior than a poorly designed cohort study. Effectively, there are many things that can only be done in case-controlled studies, especially for rare outcomes.

Now, what we've done at CDC with a lot of global partners is begin to put our finger on the pulse of the so-called world of human genome academiology. We have this database of all the literature. This is only the published literature that we've been gathering since October of 2001. Essentially there are more than 15,000 association studies that are being published from only over the last three years. Those numbers are increasing.

Most of the data come from association studies. Most of them are case-controlled studies. There is an increasing number of studies that focus on gene/gene and gene/environment interaction, and there are a few studies that are just pure prevalence of different genetic variants in populations. But this is where the action is.

We are actually doing a 5 percent random sample of this database to look at the quality of these association studies. But other people have looked at that and have found that many association studies have poor quality in terms of epidemiologic parameters.

NHANES was alluded to earlier. This is a study to look at the prevalence of the top 50 genes of public health significance that we are collaborating with NIH on to measure in the NHANES III, which is about 8,000 representative samples in the U.S. Those sort of 87 SNPs and 57 genes, and then trying to correlate those with the 2,000 phenotypic variables that already exist in the NHANES III bank.

This is another example of a population-based case-controlled study that essentially uses surveillance systems which are population based. These are surveillance systems for birth defects that are doing case-controlled studies for looking at genes and environments in relation to birth defects. There are about 10,000 cases and controls, and those numbers are going up.

If you have a population under surveillance like you have, it is equivalent to a cohort study of more than 1 million persons, or 1 million births, at least. There are other situations where you can do either massive case-controlled studies, or cohort studies like in managed care organizations.

So why do we need to integrate data? We have unmanageable amounts of data, two genes, three genes, four genes. For most chronic diseases, common diseases, we are at least dealing with 10 to 15 genes to explain most of the etiology.

We have small sample sizes, whether we look at cohort or case-controlled studies. I'll show you a slide on that. We have small expected effect size of gene disease associations. Why? Because most genes are not expected to contribute by themselves to the etiology of most of these diseases. So the rule, rather than the exception, is to expect relative risks or odds ratios that are close to 1.3 or 1.4. So you need large sample sizes to discover them.

You need replication across studies. There is a lot that we have been dealing with with publication bias. There is heterogeneity that we have across populations and within populations, and you need to both generate and test hypotheses.

This is data from John Ioannidis from Greece, who is part of the HuGE movement, and has been really keeping his finger on the pulse of the published association studies. Most of these are small sample size, probably 200 or less. Most of the hundreds of gene disease associations have odds ratios between 1.0 and 1.4. This is sort of the peak at 1.2.

So how do we build the knowledge base on genes and population health? The answer here is all of the above. But let me go through this thing with you. Single large population cohort study, a systematic synthesis of data from existing and planned cohort studies, a systematic synthesis of all data from either cohort studies, case- control, or all of them. The approach we're doing is number four, which is an accelerated systematic synthesis of both group and individual data using collaborative networks and consortia of all types of studies.

Of course, the right answer is number five here. But what do I mean by that? In 1998, CDC and many partners developed the Human Genome Epidemiology Network, which is truly a global, open-ended collaboration of both individuals and organizations that are interested in assessing the population impact of genomics on health, and how we can use genetic information to improve health and prevent disease.

The network has about 700 people right now from 40 different countries. It is wide open to anyone who wants to join it. There is a website with information exchange. There has been a lot of training and technical assistance through the form of workshops that we've been doing. Roughly on average, one a year.

We are developing the knowledge base, putting stuff together in terms of synthesis with quantitative methods of matter analysis, and we want to disseminate information for policy and practice.

You have already seen the huge studies database that I alluded to earlier. In addition to that, we have been sponsoring in collaboration with six journals, systematic reviews of gene disease associations that many authors have subscribed to. We also have a database of 200 meta-analyses of different gene disease associations that is published elsewhere.

I mentioned the methodology workshops. I'll mention briefly the international biobank cohort study meeting we just had. We are in the process of forming a network of 14 different networks that exist in the world. Many of them are in cancer. Some of them are in heart disease. These are networks of investigators that have come together to pool their data and share information.

We are developing the sort of sharing of information between networks. Just by the way of going through this whole cycle from funding to publication, very quickly going through where things are right now. We are talking about different study designs, whether it is biobanks in one study, case-controlled studies or consortia, people do these studies, and then they report them. Then somebody else will appraise that literature, review it in the form of meta-analysis, cover methodologic problems and research, and then the funding cycle continues.

What HuGE Net is trying to do is influence the circle here. We are collaborating with the various biobanks. We have focused primarily on this region here, but this will influence the study designs as well. I don't have time to go through this.

This is courtesy of Marta Gwen from our office that has superimposed this on an elephant, because depending on where you are in the world and what kind of studies you do, you only see part of the elephant. What HuGE Net is trying to do is to look at the whole elephant together.

This is briefly the meeting we just had in Atlanta in collaboration with P3G and NIH, courtesy of Teri Manolio. We brought together a small group that talks about the harmonization of epidemiologic data. This is the outcome of this meeting.

One of the outcomes was, and we are working on it, a statement that would be essentially important for publishing studies that are derived from biobanks. You might say well, the data won't be coming until 50 years from now. But if you have a statement, it refers to a movement in the world called Standards for Observation Studies in Epidemiology. This is a worldwide movement. U.K., Canada, and the U.S. have been setting standards for epidemiologic studies outside genetics. What we are trying to do is influence the conduct of biobank projects and biobank studies through developing similar criteria.

The biobanks themselves are going to put together sort of best practices for the design and conducts of biobanks, and then update their online knowledge base with a register of studies and tools, and then having further meetings.

So in conclusion, these are my three messages for today. One cohort study in one country is not enough. There is more than one way to get there. I think all the ways will get us there. What we need to do is work all together to really look at this challenging area ahead of us, which is how do we make sense of the Human Genome Project.

Thank you.

DR. TUCKSON: Thank you very much, Muin. I appreciate it.