Overview of International and National Large Population Studies
Teri Manolio, M.D., Ph.D.

DR. TUCKSON: Now Teri Manolio will give us a sense of the overview of this issue from the international and national perspective. Thank you so much, Teri.

DR. MANOLIO: Great. Thank you very much.

I appreciate being invited to comment on international and national cohort studies. There are a large number of them and we won't be able to do them all justice. Luckily, several will be discussed in more detail here.

So what I was asked to do was to review these studies and then talk somewhat more about design as well, design of prospective studies versus case-control studies, design of phenotypic definition, and I probably won't have a chance to get to this last one, use of existing cohorts versus new cohorts, but if we time we'll do that as well.

There are, as I said, a large number of these. There are new ones sort of cropping up every day. Very few of them had actually gotten into the field and gotten going.

The Public Population Project and U.K. Biobank you'll hear about a little more from subsequent speakers, so I won't focus as much on them. Biobank Japan and Estonia I can talk about a bit, and this one I can go into a little bit more detail because it's actually the one that's furthest along and is generating results. I'll also comment on the Marshfield Project, you'll hear about the National Children's Study, and there are a variety of other clinical samples that I won't go into.

Just a broad overview of several of the international ones, the Biobank Japan, obviously in Japan, is anticipated to be 300,000 people ages 20 and above. The focus at present is on 47 common complex diseases, which, as we've heard before, were diseases that do not seem to have Mendelian patterns of inheritance that are related to a single gene, but probably to multiple genes. Access to those data and samples at present is limited to Japan and Japanese researchers.

DeCODE Genetics was mentioned earlier. It's in Iceland. They anticipate having most likely the entire population if they keep going, at least all of those that consent, which would be at least 200,000 of all ages, 50 common diseases, and access is possible with collaboration.

The Estonian Genome Project in Estonia has varying estimates of the size. The total size of the country is about 1.3 million and they had initially talked about trying to get a million of those. Now they're scaling back a bit more to closer to 100,000. The age I'm not quite sure of. I assume it's all the adults, but I don't know. Common diseases, and again with collaboration.

Then you've heard much about U.K. Biobank and we'll hear much more about that.

CARTaGENE is a Canadian study in Quebec. It's anticipated to be about 50,000 people aged 25 to 74. Again, focusing on common diseases, and Mylene, who will be filling in for Bartha Knoppers, whose flight was canceled, will tell you more about that perhaps.

GenomeEUtwin, similarly, is part of that collaboration. It has seven European countries with 800,00 twin pairs. Twin pairs are a very interesting genetic model. They have great strengths, as

well as some weaknesses, and I'm sure you'll hear about that. It's focusing on seven key outcomes at present, and they are available with collaboration.

The Marshfield Personalized Medicine Project is in Marshfield, Wisconsin, relying on the Marshfield Clinic. It anticipates 40,000 people 18 and above with a very large focus on adverse drug reactions. David Goldstein spoke to you earlier about the importance of adverse drug reactions, and I think that would be a place, David, where you could find some really exciting information about this.

The National Children's Study Dr. Brenner will be talking about a little bit later. It's to include 100,000 infants and their mothers and to follow them for 21 years.

Just briefly to comment on Biobank Japan, the goal of the study is to clarify on a large basis the causes of diseases and medication side effects in relation to genetic variations and ultimately to develop new drugs and diagnostics.

The goal of many of these large biobanks is focusing towards drugs and diagnostics as a way not only to contribute to the field, but also to help support the biobank itself.

Samples and data will be collected and are being collected by a network of collaborating organizations and private universities. Public universities are not involved in this one, and that has raised some eyebrows, as it were, outside of Japan, but the Japanese seem quite happy with it and it's their study.

These are some of the universities that are involved. The Tokushukai group bills itself as the "third largest hospital group in the world," and it does have a very large catchment area.

They hope that their project will stimulate the development of legislation in Japan to protect personal research information. Not only genetic information, but research information in general, which is an interesting sidelight to the biobank.

It was begun in 2003. Ninety-thousand samples have been collected to date, and that actually is 120,000 disease cases because each person that they've collected has more than one disease. This is unlikely to be a random population sample. It's more patient-based because it's working with hospitals, and so its relevance to a general population is a little more questionable.

Distribution of DNA and serum to Japanese researchers has already begun.

The Estonian project has a similar goal to find links between genes, environmental factors, and common diseases, and apply that to improved health care. There may be as many as a million persons, but now scaling down perhaps to 100,000, and it was begun in October of 2002 with about 10,000 recruited in an initial pilot as of 2004 in three Estonian counties.

There is written informed consent, a 60 to 90-minute questionnaire that includes genealogic information at least back two or three generations, simple measures -- height, weight, blood pressure, heart rate -- and a 50-milliliter blood sample.

Personalized information is intended to be provided back to participants with their consent and with their interest, and to their physicians, again with their consent. The people who participant in this are called "gene donors," and actually participants can go on to their website in Estonia and ask a series of questions about their involvement and what it means for them.

There is a non-profit Estonian Genome Project Foundation which is in public/private partnership with eGene, Inc., which was a private arm. Actually, they have just recently dissolved their arrangement with eGene in 2004 and they're now looking for other sources of funding.

The Marshfield Project, as I mentioned, is based out of the Marshfield Clinic in Wisconsin, which is a very large private set of clinics. It's intended also to translate genetic data into knowledge that will enhance patient care.

It utilizes the Marshfield Epidemiologic Study Area in Central Wisconsin, which has a longstanding electronic medical record program, and so utilizes the strength of having ongoing electronic records. I would comment, though, that clinicians are still clinicians, even in Wisconsin, and they don't always record things in a standardized way. So just because it's electronic doesn't mean that it's reliable.

There are active programs in Marshfield in genomics and clinical research. They intend to recruit up to 40,000 people aged 18 and older. This was begun in September of 2002 and 17,000 recruited so far. Response rate is actually fairly respectable for a study of this size and scope, 45 percent. In epidemiological studies, we like it to be much higher, but for a variety of reasons, this is quite good.

There is written informed consent, a 30-minute visit with questionnaires, DNA extraction, blood. The data are encrypted, which means that there is no one with access to the identifiable clinic information has also access to the genetic information, and there's a link there that can be broken by a third party.

DeCODE Genetics is the Icelandic group. They are a biopharmaceutical company that are applying discoveries in genetics to develop of drugs for common diseases.

They utilize the unique resources of Iceland, which is that, first, it's relatively isolated. It's an island in the middle of the North Atlantic. There are founder effects there, which means that they were settled by a relatively small number of people -- probably in the tens of thousands, though, still -- in the early 10th Century, and it remained isolated since then. They've also gone through a series of population bottlenecks, famine, disease, and volcano eruptions and things.

They also have an extensive genealogic database extending back to the settlement of the island in 900 A.D. They have a very small number of high quality referral hospitals and very good records.

DeCODE currently has DNA and data on 110,000 consenting Icelanders and about 25,000 non-Icelanders from various parts of Europe that they have collaborations with. It was begun in 1998.

There was tremendous controversy generated by this project, primarily because of their proposal for an opt-out consent for access to medical records. There was a proposal to have what was called a health sector database that would be accessed in everyone, and this opt-out consent did cause a big problem. That eventually was abandoned. The plans for that, whether they'll be revisited or not in Iceland is not clear, but there has been written informed consent for all of the genetic studies, and there's third-party encryption as well.

I should, in the interest of full disclosure, mention that I am collaborating with this group. So that's partly how I know a little bit more about it, but you may want to take my comments in that context.

The uniqueness of this population, as I mentioned, they were founded by settlers of mixed Northern European descent from Norway and Sweden. They stopped off in the British Isles and picked up some passengers, sometimes willing and sometimes not, and went to Iceland from there.

The current population is about 285,000, which is almost exactly one one-thousandth of the U.S. It's about the size of the town of Framingham, which you may have heard of, and another tremendous resource is their careful genealogic records. Genealogy in this country is more than a national hobby. It's almost an obsession. I mean, they all know who they're related to. When two Icelanders meet, they'll say, "Oh, you're so and so's grandson. My cousin went to school with your aunt," and they can all relate each other to various and sundry relatives, and without any enmity or anything. It's not like there are feuds between clans and that sort of thing, but it's clearly something that they're very interested in and have kept very good records.

So given the relatively small founder population, there is relatively similar genetic background, and their isolation following that means that there are fewer variants to study.

What has been done with these genealogic records -- which any family, if you visit an Icelandic home, they have books in their family and after dinner they'll take them out and show you how they relate back to various groups -- is these have been computerized, and every Icelander has a password to this.

This is actually the genealogy of Kar Steffenson, who is the founder of deCODE, and he can go into this, as can any Icelander, and trace his genealogy back one, two, three, four, five, six generations to this person. Then click on this next button, and she was born in 1776, and trace her back another six generations. Then the next one, born in the 16th Century, and in the 14th Century, and in the 12th Century, and finally back into the 10th Century. So back to their original Norwegian founders. Most of them can do this. It's really quite remarkable.

What they also can do is when they meet someone, they can go home and look them up in this database --

(Laughter.)

DR. MANOLIO: -- and found out who they're related to and find out how closely they're related to each other. So married couples, it was very interesting when this came out. They were saying, "Oh, we're actually related back five or six generations. Maybe that's why our son Charlie is so strange."

(Laughter.)

DR. MANOLIO: More often, it's just an interesting hobby that they have. They're very interested in it. They'll say, "Oh, I can go home and check and see who I'm related to," and this is a big deal for them, so that's fine.

It's also a big deal for science because what one can do then is take two people that happen to have the same disease and see how they're related to each other and pull out groups of cases that actually are related in very large pedigrees.

That was done in our atrial fibrillation project. This is a pedigree with 69 patients. It's not the largest one that they had. There was one that was 700, but this one fit on the page.

What this shows you is that all these people with atrial fibrillation in these little black boxes and circles, which are a tremendous resource then for finding genes, and the purpose of this kind of study is to actually identify genes related to common diseases.

What we did with this then, recognizing that common diseases don't show Mendelian inheritance patterns and very often you don't just have affected sibs, which is the model that's most often used in this country looking at sib pairs, but you often have people with more distant relatives. So you can look at the degree of relatives.

If you have a person with atrial fibrillation, his or her first-degree relatives are 77 percent more likely to have atrial fibrillation than people without a relative with atrial fibrillation. If you exclude the first-degree relatives, which are mothers, fathers, sisters, brothers, daughters, and sons, the relative risk is still 36 percent higher, 18 percent higher if you look at third-degree relatives, 10 percent, and 5 percent if you look at fifth degree.

Very few populations can go to this level of detail in relationships, and what's interesting about this particular example is that this decline by halves basically in degree of relative risk parallels the decline in sharing of genetic variants through generations. So it's a very strong suggestion that there's something genetic here that is related to this disease.

So deCODE has used this approach to map diseases, which means finding areas of chromosomes that are likely to be related to disease for all of these diseases shown in white here. For those shown in blue, they've actually identified what likes to be a causative variant. So within a gene, they've found the gene and the possibility of a variant related to it. Then these purple ones are things that they've actually developed drugs for and are in clinical trials to try to reduce. So again, a very powerful way for finding genetic variants.

Now, one of the challenges in identifying genes is to actually understand, as Gil was alluding to earlier, the population impact of these, and I guess I would quibble a bit with Dr. Goldstein's comment that just because you know a gene, you can't do anything about it.

ApoE4, for example, we actually know interacts with a variety of other risk factors in relationship to cognitive decline, and it may be that one would want to really reduce those other risk factors as a way of perhaps reducing the risk in someone with ApoE4. That's a reasonable research question that needs to be pursued.

But if you consider genes just to be risk factors passed from parents to children, epidemiologists know what to do with risk factors. Then you want to determine the prevalence of them. You want to look at associations that are identified in family studies or other studies, and assess their magnitude and independence, recognizing that common risk factors are generally not strong ones and strong risk factors are generally not common. If they were, we'd all have them and we'd all be sick. So basically, those get weeded out and we end with the smaller effect, but that are much more common.

One can define associations with a variety of phenotypes. Not just atrial fibrillation, but perhaps as it's related to other diseases as well, and identify factors, particularly environmental factors, because these are the things that we can change. These are the things that have changed in the past 30 years to give us this incredible epidemic of obesity that we're facing. That hasn't been the genome that changed. If we can identify those things and have some impact on them, we may particularly want to do that within genetically susceptible individuals.

This shows just three of the variants that deCODE has identified. There is a little bit known on the allele frequency and the risk associated with these in the Icelandic population. The Icelandic population, for a variety of reasons, is very different from the U.S. population, and one would want to know not only the allele frequency and the risk, but other phenotypes and associations are there with these particular variants? And particularly, what modifies them? Very little of that work has been done and that's what needs to be done in these larger biobanks.

Francis Collins published a paper earlier this year talking about the need for large cohort studies, and Dr. Guttmacher will comment on this a little bit later.

Identifying and reducing disease risk depends on an unbiased determination of a variety of things. The actual quantitative contribution of both the environment and the genetic factors, the interactions among them, and the interplay among other disorders that may share common risk factors. So if you get heart disease, does that affect your risk for asthma or cancer or other things? It probably does.

He recognized and pointed out that replication of associations and estimating their magnitude, consistency, and their time relationships is best done through prospective cohort studies.

Just briefly, cohort studies are prospective -- that is, from before the time a disease develops out into the future -- investigations of a representative sample, representative meaning that you can relate that back to the population from which it was drawn. So you're not just studying truck drivers who may be different from the rest of the population. You're not just studying Air Force pilots. You're taking a sample that's representative of the entire group.

You follow them for development of specified endpoints. So you want to identify things and look for them actively, so that they don't just happen to be picked up, but actually are surveyed and picked up systematically.

The purpose, as mentioned before, is to identify risk factors predisposing to development of the disease in general populations. Particularly, you want this design when you're looking for risk factors that are affected by disease. So you can't measure them after the disease has occurred, the things that are affected by treatment or by lifestyle changes. When people feel sick, they might think I need to do something about it to prevent myself from getting disease, and so those things can then have an impact on the associations you measure.

You particularly want to look at those that are difficult to recall or in which there is biased recall once somebody develops a disease, and we'll talk about that in a minute, or with hypothesized early pathogenic effect. So something that has an impact early on and then later on may not have much an effect at all, you're likely only to pick those up in prospective studies, rather than waiting until the disease occurs.

And they complement a variety of other epidemiologic designs which I'll talk about, particularly case-control studies.

Again, in the interest of full disclosure, I should mention that I'm responsible for the group at the Heart, Lung, and Blood Institute that runs major cohort studies, such as Framingham, Honolulu, and a variety of others. The sample sizes are shown here and the ages, and fortunately we're doing a little bit better in including minorities, but that has been a challenge.

Pros and cons of these kinds of studies. They are very expensive, they take a very long time, you need large numbers of people, and they're very broad-based, and so there tends to be a lot of criticism of them as being fishing expeditions, et cetera, et cetera.

They, however, provide risk information that really you can't get any other way. Healthy people don't typically go to the doctor, and they don't get screened and they don't get their risk factors measured, and if you want to understand why healthy people get sick, rather than why sick people get sicker, what you need to do is a prospective study.

In general, the public is better able to understand these than often with clinical studies because you can relate to the people. "Gee, that's somebody just like me. That isn't somebody that was exposed to beryllium," or whatever it might be. "It's somebody just me living in a community. I can understand that."

They identify modifiable risk factors that might be intervened upon, which is what we're in this business for anyway.

If you wanted to look at the characteristics of ideal cohort studies, size is very important. The larger, the better, up to some degree, obviously, because when they get to be too big you may not be able to actually measure enough on them to make them worthwhile.

They should be representative. They should be diverse in geography, in this country, at least, socioeconomic status, and race/ethnicity.

There should be standardized and reproducible characterization of exposures and risk factors. Ideally, there should be repeated interim measures to check differences or changes in risk factors and exposures over time, and comprehensive standardized assessments of outcomes.

If one doesn't do this, particularly the standardized aspects of it, you're prone to a variety of biases that can affect your study results and lead to basically erroneous conclusions. I've mentioned a number of them here. Several of these are particular problems in the case-control study design, and case-control studies have gotten a bad name mainly because I think people haven't followed appropriate design strategies for them.

These are three assumptions that one has to basically meet in order to have a well-done case-control study. The cases are representative of everybody who developed the disease. Not just the people who go to Hopkins, not just the people who drop dead, but everybody.

Controls are representative of the general population that don't develop the disease.

Most importantly, collection of risk factor and exposure information is the same for cases and controls. This can be a real problem because once somebody is sick, it affects the way they recall things and the way they report them.

The advantages of this are it may be the only way to study rare diseases.

Existing records can often be used if the risk factor data are collected independent of disease status, and that often doesn't happen. Once somebody has lung cancer, you ask them 1,000 times if they smoked and were exposed to asbestos and that sort of thing.

You can study lots of etiologic factors, and they may be less time consuming and expensive.

Disadvantages are that they rely on recall or records for information, and validation of these past records can be very, very difficult. Selecting an appropriate comparison group can be tough, multiple biases, as we talked about before, can get spurious evidence of associations, it's difficult to study rare exposures, and it's difficult to study temporal relationships.

Now, it's usually at about this point in a conversation with geneticists that they say me, "Now, wait a minute. This is genetics, you dumb epidemiologist. This is different. Genes are measured the same way in cases and controls. No bias there." Information on your key exposure of the genes, then, is very easy to validate. There's no recall or reporting and temporal relationships are very clear.

But in response, I would say that bias-free ascertainment of cases and controls is still a major concern. Cases in most clinical series are very unlikely to be representative and assessment of risk modifiers or gene/environment interactions is very likely incomplete or flawed unless you have done it in a prospective way.

But this is a very, very powerful design. If you look at a disease with an incidence of 8 per 1,000 among the unexposed, which is a relatively rare disease, a cohort study would require 4,000 exposed and 4,000 unexposed people to detect a two-fold increase in risk. A case-control study would require only 200 cases and 200 controls with a 30 percent exposure. If you then look at disease that's a quarter as common, 2 cases per 1,000, you need 16,000 exposed and 16,000 unexposed to detect that same degree of risk, but a case-control study still requires only 200 cases and 200 controls.

So this is a very powerful design, and what to do, and I'll finish up in just a moment, is to nest this kind of study within a prospective study, so that you identify cases as they develop and them measure on them things that would otherwise be very expensive to measure in an entire cohort, because a large proportion of the cohort members never get sick and they don't contribute very much incremental information. So if you can collect information and store it, as in blood, as in DNA, et cetera, you're able then to apply this design, and you can expand it to other types of study concepts.

I think I'll stop here at this point and see if there are questions and go from there.

DR. TUCKSON: Well, thank you very much. Very, very good.

Any hot questions right now? If not, we'll come back.

(No response.)

DR. TUCKSON: Well, thank you for that.