[MALE RESPONDENT]
[Other comments:]

I'd like to welcome to the stage, Professor Sek Kathiresan, Director for the Center for Human Genetic Research at the Massachusetts General Hospital, and Co-director of the medical and population genetics program at the Broad Institute of Harvard and MIT. [?Sek 0:00:18.9] will share some of his experiences of working with the existing genotyping data set at UK Biobank. Thank you Sek.

Okay, thank you very much Mark, really a pleasure to be here to speak about our experience using the UK Biobank, what I consider to be the world gold-standard in terms of an open data resource for health research. So, the key public health need, of course, is to identify individuals at high risk for disease. Since most diseases have an inherited component there's an idea to use DNA variation, inherited DNA variation to stratify individuals at risk for disease. Historically, if you take a disease like myocardial infarction or heart attack, the problem that we've studied, monogenic mutations have been the basis for stratifying individuals. A little bit of background here in terms of myocardial infarction it's a complex trait, basically involves two phases, a chronic phase where there's build-up of plaque in the heart arteries and then there's an acute phase of plaque rupture [?inside to 0:01:32.2] thrombosis and necrosis of heart tissue that can be detected by symptoms, EKG change. or elevation and cardiac biomarkers.

Like many diseases it has an inherited component and lifestyle component. It's a complex trait and for about half of all MI as the first presentation can be sudden death. If you think about the inherited component, again monogenic is really what's been focused on before. What I'm going to tell you today about is the polygenic component. The monogenic component refers to, of course, a single mutation sufficient to lead to early disease. Polygenic refers to the additive effect of many variants, in aggregate can push somebody to disease at a young age. We focus specifically on disease at a young age because inheritance plays a larger role. In terms of the monogenic component if you think about MI, gene variants that increase LDL cholesterol in a couple of different genes in the genome, are prevalent in about 0.4 per cent of the population and they increase the risk of heart attack by about three-fold. These variants work by increasing the level of LDL cholesterol in the blood.

Now, again, the question we set out to address is beyond the monogenic model, can we identify additional patients with high risk, based on the polygenic model for risk? The concept here is similar to the earlier study on the Type 1 diabetes, is polygenic risk scores. If you have N number of polymorphisms in the genome that are related to disease risk, and each individual can carry zero, one or two copies of the risk [?allele 0:03:11.5] and, so, a score can range from zero to two N for that person. This score can be unweighted or weighted based on each variant's effect size. Over the years, we and others have constructed genome-wide, our constructed polygenic scores to really predict risk for coronary artery disease. Typically, these scores have involved gene variants that are just above this genome-wide threshold, so, at the top of the list of any association results file. There was an insight a few years ago that actually there's much more genetic

information in the entire results, the entire set of genome polymorphic features, polymorphisms in the genome and that you could potentially do a better prediction by incorporating, by using not just the top snips but really the entire feature set.

We moved from using the top snips in the genome to really a genome-wide set of almost 6.6 million variants for prediction of coronary artery disease. This is work led by [?Ahmed Kara 0:04:15.6] here in the group. The hypothesis here is that a polygenic score including a genome-wide set of snips can identify individuals at risk for MI equivalent to that of the monogenic mutations that I mentioned earlier. There are three steps here in this study design. First, is a training data set that really is a genome-wide association study that we published a couple of years ago that involves about 60,000 cases, 120,000 controls. At the end of that there are about 6.6 million variants with each variant basically having an effect size for coronary artery disease. That's really the training dataset. The second step is to develop polygenic scores using this training data set and we constructed about 24 different scores and tested each of them in the first phase of UK Biobank which was 125,000 individuals, about 4,000 of them cases, 120,000 of them controls. All of the genetic data really came from the genotypes, both the directly genotype as well as the imputed genotypes from the data set that was released last year, that Mark just mentioned.

We constructed the 24 scores, we picked the best one based on this validation data set, and then took that to really the second phase of UK Biobank which was 300,000 individuals again with genetic data. Here, there are about 9,000 individuals that had coronary artery disease and about 288,000 controls. This score, the best score that we picked from the validation data set really worked nearly identically in the testing data set. There was really no overfitting that we saw. The data I'm going to show you going forward is from this testing data set. Here is the score for each individual across all 300,000 individuals in UK Biobank and standardized to a mean of zero, and a standard deviation of one. The x-axis is a polygenic score and the y-axis is a density plot and what you can see is this number, basically a quantitative metric of one's liability to heart attack.

This number has a normal distribution, a beautiful bell curve in the population like many other traits, like LDL cholesterol or height. This is really, again, a new quantitative metric of one's genetic liability to MI. A couple of questions come to mind after seeing this number. One, is how does it compare to the things that we use in daily practice right now to predict risk of heart attack? On the x-axis here, is the polygenic score for a set of individuals. On the y-axis is something called the American College of Cardiology, American Heart Association Pooled Cohorts Equation which is the risk engine that we calculate in the US to predict risk for MI that incorporates age, blood pressure, cholesterol and so forth. What you can see is the correlation between the polygenic score and the existing risk score is very low, suggesting that this genetic information is orthogonal to the things that we measure every day in clinical practice.

Getting back to this question using this polygenic model, can we identify a set of individuals with MI risk, equivalent to monogenic mutations? Here on the x-axis our percentile bin, one per cent is each dot of polygenic scores. Each dot represents about 3,000 individuals of the 300,000 in the test data set. On the y-axis, is the prevalence of coronary artery disease in each bin, and what you can see is there's more than a 20-fold

gradient across all these bins, and with some bins here with much greater prevalence of coronary artery disease. On this side, on the low end, there are some individuals that are very protected from coronary artery disease based on their genetic background. We ended up labelling the top five per cent of this distribution as high polygenic risk, and asked what their risk is compared to the remainder of the distribution.

We did this, we're essentially dichotomizing a quantitative trait mainly to make an apples-to-apples comparison to the monogenic mutation analysis which typically compares a carrier of a mutation with the non-carrier, in terms of risk. When we do this, we find that the top five per cent of the distribution is about three-fold increased risk, so really a level equivalent to the monogenic mutation analysis that I showed earlier. If you go a little bit further out in the tail of the distribution, so the top one per cent, for example, compared to the remaining 99 per cent, you have almost five-fold elevation and risk. In terms of comparing the monogenic and polygenic models, the prevalence of the monogenic mutations is about 0.4 per cent here for polygenic high risk. This cut off is five per cent. The odds ratios are similar for coronary disease. For the monogenic mutations these individuals who are at risk can be identified based on sequencing, but they can also be identified based on their LDL cholesterol level in the blood.

An important feature of the polygenic risk is these individuals are currently unaware. They actually do not have particularly high LDL cholesterol levels, or blood pressure levels, or other risk factors. You really do need the genetic data to figure out who these individuals are. The mechanism of risk for monogenic we know it's LDL. For the polygenic, for lack of a better term it truly is a [unclear word 0:09:51.6] of multiple things, many, many things that is leading to people to this extreme tail of the distribution. Now, in terms of when you think about genetic risk we often want to identify individuals at risk, but also more importantly be able to offer an intervention to modify that risk. For monogenic we know that lifestyle but, more importantly, medicines that lower LDL cholesterol make a big difference. Open question was, over the years, for polygenic risk is that risk modifiable? Similar to the monogenic. We've addressed this in the last couple of years and we've shown that, in fact, the DNA is not destiny here, that the polygenic risk is modifiable first by lifestyle.

We've shown that if you lead a favourable lifestyle, you can cut the risk and the inherited risk, by about half. Then, also statin medications can lower the risk that comes with the polygenic model. To summarize then, if you think about the monogenic versus polygenic contributions to MI, particularly early MI, if you have 100 individuals with a heart attack at a young age, roughly two of them will have a single mutation that leads to early disease. Remarkably, more like 17 or 20, or so, will have high polygenic risk. Then, there will be some individuals that actually have high polygenic risk as well as a monogenic mutation. Their risk is [?additive 0:11:20.2] basically almost six to eight-fold compared to those who don't have that. You may be wondering with this approach of a genome-wide polygenic score that I've showed you for coronary artery disease, how does that generalize or does it generalize to other diseases?

Within UK Biobank we've done the same analysis for a range of other diseases. Here's atrial fibrillation, diabetes, inflammatory bowel disease, breast cancer, and the pattern is remarkably similar for each of these diseases. Where, if you look at the one per cent bins there's always a tail of the distribution where the

risk is much higher compared to everybody else. This is a way to quantitate that, to look at the proportion of the population for each of these diseases based on the polygenic model that is more than three-fold higher risk compared to everybody else. You can see for example for breast cancer, two percent of women are three-fold increased risk compared to everybody else. These are women that, for example, would not be identified right now as being at high risk for breast cancer in the clinic. To conclude, now it's possible to score the polygenic component to any complex trait from genotyping array data, and this can be done simultaneously for many diseases and really early in life.

As I've shown those in the extremes of score are at risk for disease, approaching or exceeding monogenic mutations in some cases and I think it's now time to consider integrating these scores to guide prevention treatment and screening strategies. Thank you.

[END OF TRANSCRIPT]