Using Genomics to Understand an Individual's Risk for Common Diseases  JULY 2019

M: Thank you very much. It was a great pleasure in my academic role to be a central part of the gathering of the genetic information, the genotype data on UK Biobank. What I'm going to talk about today is work that we've done in our spin out company Genomics PLC around polygenic risk scores. Many of you will be familiar with the basic idea, but for those who aren't, we've learnt from ten or fifteen years of studies of complex human diseases that for any common human disease - indeed for almost every human trait other than Mendelian traits - there are many hundreds or thousands or maybe even tens of thousands of genetic variants across the genome which affect an individual's risk of that disease. Individually those variants have a small effect on risk. If you have an A rather than a T at this place on a chromosome, it might increase your risk by two per cent and a C rather than G here increases risk by three per cent and so on. The idea has been around for at least ten years of aggregating this information across the genome into what you might think of as a score and what has become a polygenic risk score, that measures for an individual the impact of these many thousands or tens of thousands or in some cases hundreds of thousands of genetic variants in terms of risk of a particular disease.

You can imagine if you did that across a large number of individuals in a population, most individuals would end up with scores in the middle of a range. They'd have some of these variants which bump their risk up a little bit and some of the variants which bump their risk down a little bit, but there will be some individuals who have been for that disease unlucky. They will have rather more of the variants that increase their risk and what has changed - and as Gil said, UK Biobank is central to this - what's changed in about the last 12 months is our ability to assess at scale in a cohort like UK Biobank how useful this could be in terms of understanding risk for specific diseases. I'll start by showing you a number of pictures that followed from our work. Other groups in the academic context have done very similar things.

So this picture represents polygenic risk scores for coronary disease in this case in men of European ancestry in UK Biobank, about 150,000 men who contribute to the data underlying this picture. So what we did is look at large studies of coronary disease not including individuals in UK Biobank. Within genomics we actually have data on about 10,000 GWAS' and we leverage that multi-trade information to generate polygenic risk scores. So we determine the algorithm for the score, which snips will be involved, how much do you weight each of them. We did that outside UK Biobank and then take it into UK Biobank and ask for different sets of - in this case, men - if we partition them on the basis of their polygenic risk score - what happens in terms of their trajectory and the rate at which disease develops over time?

In the picture that's shown here, the red curve represents the individual in the top three per cent of polygenic risk scores for coronary disease and it's using the health information for UK Biobank in a Kaplan-Meier way, to estimate incidents of disease at different age points. The blue curve at the middle 20 per cent and the green curve are those with the lowest risk in terms of polygenic risk score and there are a couple of things that follow from this. The first is, the red curve is much higher, so if you compare vertically, that speaks to relative risk. The relative risk of a man in this group is about four or five fold that of the average and actually if you look horizontally it says that a 45 year old man in this group or that group had the same
incidents of disease as a 55 or 60 year old man who is typical and maybe a 65 or 70 year old man in another group. I’ll come back to that point later.

This is breast cancer, so as you are very aware there are two genes where mutations of a particular kind have a major impact on a woman's risk of breast cancer, BRCA1 and BRCA2. This is ignoring those. This is again just combining information from many hundreds or thousands of genetic variants and you can do that outside UK Biobank, take it into UK Biobank, look at different groups of women based on their polygenic risk scores and again those with polygenic risk scores in the top three per cent are at substantially elevated risk relative to the average. Their lifetime risk in UK Biobank which is probably a healthier cohort than average is almost 30 per cent. Again, if you look horizontally a 45 year old woman in this group has about the same chance of having breast cancer as a typical 55 or 60 year old woman. In the UK, we offer screening via mammograms to women at age 50. I think it’s hard to look at this picture and think if the data were available, we wouldn't want to target screening at these individuals earlier and maybe later for these individuals.

So this information can help us target screening. Also, it changes the way we interpret screening. If you imagine a 60 year old woman in this group has a positive mammogram and compare that to a 60 year old woman in this group has a positive mammogram, because the baseline rate of disease is so different, the probability that it’s a false positive is different and much lower for this woman in the red group than for the woman in the green group. So not only can it change the way we target screening, but it can and should change the way we interpret screening. Similar picture for prostate cancer. This is Matt Hancock who is currently, but who knows for how long, secretary of state for health in the UK. He spoke some months ago about his own polygenic risk scores. He was interested in this as a potential tool for the future in the NHS and he wanted to understand from an individual point of view how that impacts an individual and their thinking and he spoke about some of his results. One of them was prostate cancer and this curve is the three per cent of individuals as he explained who have polygenic risk scores like his and that resulted in him being predicted to be about 50 per cent increase risk of prostate cancer.

So we already have in the health system ways of predicting a disease. For coronary disease there are a range of known risk factors; age; whether you're male or female; smoking history; family history; cholesterol level; blood pressure and so on. We use these routinely. In fact, GPs use them routinely in the UK via a tool called QRisk and there has always been an interesting question of how do the genetic risk scores, the polygenic risk scores, relate to the tools and the risk factors that we already use to predict disease. So it’s now possible again through UK Biobank to actually work that out. It turns out that the polygenic risk score for coronary disease is almost independent of risk scores based on traditional risk factor. So formerly the correlation between our polygenic risk score for coronary disease and the QRisk score for an individual is 0.02. This shows that data explicitly. So this is saying, take the set of individuals with the highest score on QRisk which is the tool used in the UK and it's a very good risk prediction tool for traditional risk factors and then when you apply genetics you separate even further.

So amongst the set of individuals currently put in the high risk bucket, if you look at genetics, those
with unhelpful genetics are actually a much higher risk than the average of those with helpful genetics. The same is true if you look at the next level of risk down and also if you look at the lowest risk individuals. So layering genetics on top of these traditional risk factors is informative and in fact the relative change, the relative risk in each of these groups is the same, effectively because the risk scores are independent. Interestingly, and maybe surprisingly, polygenic risk scores are largely independent of family history as used clinically. The correlation is 0.08 in this case. There are interesting questions about why that's true, perhaps for another time because of time.

So we've looked at the possibility of combining polygenic risk scores with traditional risk factors and this is a somewhat busy slide that tries to give a sense of that. So if we do the calculation as UK Biobank where we have the information and extrapolate to the UK population and there are assumptions there, because UK Biobank is not absolutely typical of the UK population in many ways, but nonetheless informative. So if you looked at the individuals in the UK aged between 40 and 55, you take the threshold that is currently used is your risk of coronary disease over ten years bigger than ten per cent? That triggers a recommendation or at least a discussion about statins with your GP. So in that group there would be about a million people above that risk threshold based on traditional risk factors. If you add in the genetics, two things happen. The first one, because genetics improves our ability to stratify risk, there will be more individuals above that threshold from a million it goes to 1.3 million. The second one is, they will be different individuals. There are half a million people who are above the threshold when you combine genetics with traditional risk factors who were not on the radar previously.

Similarly, there were individuals who were considered high risk just based on traditional risk factors who are less high risk when you include genetics. The basic idea, these are effectively independent. If your risk is close to the threshold but below it, based on traditional risk factors and you have unhelpful genetics, you can go up. If you're above the threshold and you have helpful genetics, you can go down. So a couple of points here. In the case of coronary disease there is an actual intervention - in fact current guidelines recommend statins for individuals with this level of risk. So this would give us a better tool to detect and offer that to individuals in question and it sort of says that there are about half a million in this age group and extrapolating for UK Biobank, about half a million people who are invisible to the system currently who meet that threshold were the right data and algorithms available.

Genetics adds at different ages but because age is such a key factor in coronary risk, the ways these numbers would change if I looked at 55 to 70 year olds would be less, but still significant. So one key point is we can identify people who are at high risk that were not currently identified. The second one is that the traditional risk factors are very age-related and partly because age is a key part of traditional risk, but also because things like cholesterol levels and blood pressure and so on tend to increase with age. What's interesting about the genetics is it effectively doesn't change through an individual's life. So we have the possibility of identifying individuals who will be high risk early in their life and thinking about doing the studies which would help us understand whether, if we intervened early in someone's 20s or 30s say with
statins, would that be helpful in stopping the development of coronary plaque of atherosclerosis rather than waiting until it's relatively well-developed and trying to stabilise it. That needs a proper study, but it's one I think that should be done.

There are a whole lot of interesting questions about how one takes this kind of information into a healthcare system in general and into the National Health Service. John talked about the early diagnosis cohort. We're involved in that and will be doing the polygenic risk scores for those five million individuals and so thinking about how to integrate that with the National Health Service is critical. Actually in this, it's always naïve I think to say that anything is easy in terms of implementation in a large healthcare system, but in this case one natural possibility is GPs already have on their desktop computers this QRisk tool into which various pieces of information about the patient get entered. One can imagine a slightly enhanced version of that tool which also has a box for polygenic risk score which can be entered if it's available and not if it's not. There will be some other risk factors that will be missing for some patients anyway.

The risk prediction tool can then do its thing using the genetic information if it's there, not using it if it's not and it comes out with a number which is a risk prediction, which the GP can interpret and feed back to the patient in exactly the way they do currently. So I'm cautiously optimistic then in that case there is actually a route into use in the National Health Service which is potentially relatively low threshold and doesn't put additional pressure on already overstretched GPs. There are issues with polygenic risk scores. All of the data I've shown you relates to individuals of European ancestry. These scores often don't perform as well in ancestries other than the ones where the original genetic studies were conducted. There are aspects of our methods which we think are helpful there but fundamentally, we in the community need much better data in those other ancestry groups, but just to give you a sense, the left hand picture is the one I showed you previously based on UK Biobank. The right hand picture is exactly the same polygenic risk score taken of individuals of non-European ancestry in UK Biobank. There are many fewer of those so there is much more uncertainty about estimating cumulative incidents of disease, but nonetheless it's cautiously encouraging and more than one might have expected. That's the same picture for breast cancer where, although the polygenic risk score is not uninformative in individuals of non-European ancestry - and these are relatively small sample sizes in UK Biobank - it doesn't perform nearly as well. So I'm optimistic and I think the large cohort John's described will be a fantastic way of moving this forward. I'm optimistic that these approaches can make a huge difference in healthcare. I think in the case of coronary disease it's pretty clear what that difference is. It's about identifying people who are at high risk that we don't currently know about. There will be opportunities and challenges across many other diseases, but I think we can look on this as the start of a new era of what's been called predictive prevention in primary, and in some cases in secondary care. Thank you very much.

[Applause]