

Thanks very much everybody for coming back in and also for all the questions and comments to the speakers. I think we've had a great morning session. Before we go on to the session that [Naveen Sitar 0:00:16.0] will chair on further kinds of assays that might be done, I just wanted to explain that UK Biobank, as you can see in many ways, is an experiment. It's an experiment on how to build a resource, how the community can build that resource and use it. But I think it's also an experiment in; how do we make it available in a way that utilises the imagination of all of the different researchers around the world? If one builds one's own resource and just works on it, it'll be worked on in a particular way. I think what we've already seen this morning is that different people approach the same data in completely different ways and do imaginatively, interestingly different things.

The idea of the next session was really; how do we expand that further? How do we democratise research by making this resource available to as many different people as possible and encourage them to believe in that? The idea was to have an early career research competition. We had about 100 researchers put in abstracts that I'd like to thank the Epidemiological team within UK Biobank for judging. You've seen the top 20 posters downstairs, but now we are going to have the top 3 presentations. The first of those, the winner of the prize was Doctor Phillip Law, from the Institute of Cancer Research, who's going to talk to us about genetic and non-genetic risk modelling for predicting colorectal cancer. I need to make sure I get the right prize! I wanted to say that this is not the kind of party where you come and everybody goes away with the same prize. In this case there is a special prize for the winner. Thank you very much.

Thank you very much, it's an immense privilege to be speaking here today. I'm going to be speaking about some current work that we're working on and how it ties in with the UK Biobank data, looking at colorectal cancer. Colorectal cancer is a relatively common cancer. It's increasing generally worldwide, particularly in developed countries. In the UK it's the third most common cancer in both males and females, worth approximately 40,000 new cases every year. Despite improvements in treatment, the overall survival of colorectal cancer is thought only about 55 per cent. However it is very stage dependent, with early diagnosis leading to a much better survival. As a result of the numerous screening programmes, such as the NHS bowel cancer screening programme, looking at identifying pre-malignant lesions so before the - they'd be developed into cancer, basically.

Typically the screening tests involve faecal blood test, moving generally now towards immunochemical tests and colonoscopies where necessary, each with their own advantages and disadvantages. Typically the screening is done on an age basis, so in England and Wales it's - people are invited to a screening once you're over 60. However screening programmes are generally best suited or best optimised when they're targeted at those individuals at the highest risk. So like most complex diseases, colorectal cancer is multifactorial. There's obviously your lifestyle factors such as smoking or poor diet and obesity, as well as

genetic factors such as high penetrance mutations like HPC, or common variation - which is what I'm going to be speaking more about now.

You've, I think, had enough of polygenic risk scores; it's been covered by previous speakers. Basically, it's just as we've heard, the sum of the number of risk alleles for each individual weighted by the effect size. So the effect size and the number of variants are selected, are determined using genome-wide association studies. In our study we focus particularly on just - we're a bit more conservative and focus just on the genome-wide significance snips, of which there are around 50. Once you calculate the PRS across the population, you get a distribution of the PRS. Then the hypothesis is; those with the highest number of risk alleles would therefore have the highest genetic risk.

In the previous analysis that we've done, we looked, comparing in a PRS-based screening versus a normal age-based screening and identified that the individuals at the top one per cent have a threefold increased risk over the population median. If you use PRS-based screening you'd identify 20 - you need to screen 26 per cent less individuals at the cost of missing out on about six per cent cases that would've been identified with screening. Those analyses have all been previously done on just summary-level data. We've just taken effect sizes and across what we had. The launch of the Biobank, you have individual-level data so there are around 4000 colorectal cancers, with a primary diagnosis of colon or rectal cancer, and 300,000 individuals which are cancer-free, all genetically Caucasian. In addition we have the lifestyle information tied into all of this, which is a great advantage, which I'll get into.

We see these figures from [inaudible 0:06:49.5]. If we compare the distribution of the PRS between those with cancer or colorectal cancer and those without, you see the expected right shift. Perhaps not the greatest discrimination but it is only 50 variants. Then when you stratify it based on percentile similar to - I've seen before, those with the highest population risk - sorry, so those with the highest polygenic risk have a higher prevalence than those with low. I mentioned the non-genetic risk factors so you can simply model those, include those in together with the genetic score. So these are the known ones which I've mentioned earlier, so for example as we expect; eating more vegetables, doing physical exercise is protective. Smoking, eating all the processed meats increases your risk.

That's all the easy stuff! Those are the known factors. What about the unknown risk factors? The Biobank has thousands of the various diet and lifestyle factors which we'd like to try identify. Currently building a model, trying to identify those and select which are the most significant risk factors that there. As additional validation we'll have access to the data from the other cohort studies such as bowel cancer hubs and breast cancer now. Ultimately the idea is to develop this algorithm to identify those individuals with the highest risk of developing colorectal cancer and then just to improve screening as a sort of precision preventative method. I just want to thank my group at the ICR, Amit and Richard, and of course all the participants and the Biobank team for providing this amazing resource. Go on Twitter if you want to follow our work and see what we get up to. Thank you very much.

Thank you very much, Phillip. We had two runners up and the first of these is Doctor Hanieh Yaghootkar from the University of Exeter Medical School, who is going to talk about combining UK Biobank genetics and MR imaging data to identify genetic factors associated with higher body fat but lower risk of diabetes. If anybody voted for Brexit I think we'll see. Look at all you - first two speakers, clearly not originally from the UK, and how much we're benefiting from them. Doctor Yaghootkar.

F: Thank you for choosing this research to be presented here today. It's a great opportunity to be here. In summary I'm interested in genetic factors that are associated with being fatter but healthier. If we look at the distribution of BMI - or body mass index - in UK Biobank population, you see that people who have higher BMIs have higher risk of type 2 diabetes, as shown by blue colour here. However you can see that there are many obese individuals, as defined by BMI of more than 30, who don't have type 2 diabetes. We think this is partly genetically defined. There are some genes identified for this, but to give you some example: what you see in the left-hand side here, this is an example of gene which is associated with higher body fat percentage and higher risk of type 2 diabetes, as we know because BMI is a major risk factor for type 2 diabetes.

However in right-hand side you can see an example of a gene which is associated with higher body fat percentage but lower risk of type 2 diabetes. This is very interesting because if you study these genetic factors, we can understand the mechanism that delay or protect people from type 2 diabetes. Now, UK Biobank made it possible to identify more of these genetic factors because based on half-a-million individuals with body fat percentage and genetic data was available. We combined it with data from studies of metabolic biomarkers. We characterised 14 genetic factors that are associated with higher body fat percentage, but higher levels of good biomarkers like HDL cholesterol, which is a good lipid, and also lower levels of bad biomarkers like triglycerides.

In contrast, again you can see an example of gene which is associated with higher body fat percentage but an adverse metabolic outcome. We have studied these 14 genetic factors together in UK Biobank. They are associated with higher body fat percentage but with lower risk of type 2 diabetes, and also with lower risk of heart disease and lower risk of hypertension and blood pressure measures. What's the mechanism? You might think that maybe it's a favourable body fat distribution. Again we use UK Biobank data because we had waist circumference and hip circumference. We could analyse men and women separately because they are different. As you see, these genetic factors are associated with a favourable body shape in women because they are associated with lower waist circumference and higher hip circumference. However in men they are not associated with a favourable body shape, as they are associated with both higher waist circumference and hip circumference.

Again UK Biobank, an amazing data, having access to MRI scans of body fat so we had more than 5000 individuals in the first wave of the data, with subcutaneous fat. Now, this is a safe place to store fat and also visceral fat. Now, this is a wrong place to have fat, or it's a bad fat, and also measures of liver fat. Again this is bad fat and we also added data from other cohorts to increase our sample size. So you see in these plots

here, the vertical lines are the [?null 0:13:58.1] effects. You see that these 14 genetic factors were together associated with higher levels of fat in the right place. These diamonds are the effect when we use all these studies combined. Also higher levels of subcutaneous fat in women and men, but they weren't associated with visceral fat, but were associated with lower levels of fat in the liver. This could suggest that liver fat is maybe more important in the mechanism of type 2 diabetes. Another thing was that the effect on lower liver fat was only detectable in women.

Now, these things I have shown so far is using European individuals from UK Biobank, but UK Biobank is a huge data. So we can also study other ethnic groups like for example South Asians. These are very interesting because the prevalence of type 2 diabetes is much higher in South Asians. We have data on more than 10,000 individuals of South Asians; you can see in blue colours in these plots. In comparison to European individuals, in red colours, you can see these genetic factors are associated with higher body fat percentage, higher BMI but lower risk of type 2 diabetes and lower risk of heart disease. This protective effect is - does exist in South Asians. However when we compare the distribution of these genes in the two populations, you can see the protective genes are less common in South Asians, which could explain why the prevalence of type 2 diabetes is higher in this population.

In summary there are some genes which affect where we can put our fat. We think that these genes can protect from type 2 diabetes by redistributing fat from a wrong place to a safe place; from inside to outside. Now, if genes can do this, there could be a potential for drugs to mimic the same thing. As an example and proof of principle, TZD drugs are used in clinic to treat people with type 2 diabetes which have the same mechanism. I would like to finish by thanking the Exeter team, led by Professor Tim Frayling, and also all my collaborators for all their help, support and data especially Professor Jimmy Bell, who leads the MRI imaging of body fat, also Diabetes UK for supporting my work, and also fantastic data from UK Biobank. Thank you very much.

As I mentioned, we had 100 people submitted to this from all around the world, not just from the UK; from Europe, from North America or Australasia and beyond. As you saw, the first speaker from South Africa originally, but has been here for ten years, great gain for the UK. Hanie, originally from Iran, again showing how Britain is benefiting from intellectual immigration, so fantastic talks. I'd made the case that we were going for kind of heterogeneity and research being done all over the world. But the next speaker will emphasise how people who get UK Biobank, institutions that get UK Biobank and really work on it can benefit as well. With slight embarrassment, we have another speaker from Exeter showing how they can really utilise this resource. The second runner up is Doctor Samuel Jones, who's going to talk genetic studies of accelerometer-based asleep measures in 85,000 individuals. New insights into human sleep behaviour. Sam. You're from Exeter?

I am, yes.

Good - and before I forget...

I was going to ask, but thank you! Thanks, Rory, and thank you to the committee for considering my work prize-worthy. Yes, my current work is focused on sleep data as derived from the accelerometer recordings in the UK Biobank cohorts. How did I originally get into this line of research? Well, I work in a team, as does Hanie, that focuses on metabolic disorders and metabolic phenotypes such as type 2 diabetes and obesity. We really got interested in sleep through the links with these phenotypes. Now, poor sleep and sleeping at the wrong time seems to be quite strongly associated with poor metabolic outcomes. There's a whole raft of evidence and a whole raft of literature out there to support this.

When we looked into it we realised that there aren't many genetic variants associated with sleep metrics and sleep disorders. So when UK Biobank came along and we realised that there were some questions about sleep patterns, we started working on genetic analyses of some of these measures. In 2016 we published our first paper on the genetics of whether you're a morning or evening person. We found 16 genetic variants that were associated with that phenotype. Since we've gone on to work with teams in the US and in other areas of the world, looking at the other self-report phenotypes. So far we've identified several hundred variants associated with these phenotypes.

The problem with self-report data, it's really good data but it's not particularly detailed when it comes to sleep. When we realised the accelerometer data was continuously recorded over a period of seven days in 100,000 individuals, we thought this would be an ideal opportunity to try and see if we can extract some measures of sleep. This is the activity track of an individual participant of the - who took part in the accelerometer sub-study. I've split the three axes up here. You can clearly see the periods of sleep here, but doing this algorithmically is actually quite a difficult task. So we worked with the author of a piece of software called GGIR. This software was designed to extract measures of physical activity and sleep from accelerometer data - which is perfect - but it needed some modification to work with the UK Biobank accelerometer data. Since we started working with him, he's gone on to produce a new algorithm which can identify periods of sleep without the use of sleep diaries, which the UK Biobank doesn't have.

I'm just going to zoom in on a particular day. This is a schematic illustration of the eight sleep measures that we've managed to extract using this piece of software. You'll only see several of them on here; the eighth one is about variability across multiple days. We have three measures of timing so these are more your circadian phenotypes. We have three measures of quantity, so it's about how much you sleep or rest and how variable your sleep is. We have two measures of sleep quality, so how fragmented your sleep is and how well you sleep when you're in bed. Just to give an example of the average participant in the accelerometer sub-study, so the average participant would actually sleep around seven hours 20 minutes per night. Their minimum points of activity were about 3:20 in the morning, which roughly corresponds to the middle of their sleep period. The last phenotype here is quite interesting; this is how much people move during a night. The

average participant would move between 16 and 17 times a night. That's typically shifting position, but still asleep.

We generated these measures and then we combined it with the genetic data recently released, the second release of the genetic data, to try and identify regions of the genome that are associated with our sleep measures. I apologise for the busy slide, especially for those of you at the back, but the main points here are; we identified 40 regions of the genome associated with our sleep measures, some of these overlapping in the different sleep measures. Of interest is, we picked up genes that have previously been associated with restless leg syndrome from a 2017 GWAS. We've identified genes that have been previously linked to circadian phenotypes such as chronotype. We've also identified genes which have been linked to sleep duration as well from self-report studies. We've also picked up a lot of genes that haven't been identified as being associated with any other sleep phenotypes at all. We really think we're picking up new genetics here.

Of course just to give you the bog-standard heritability estimates, they range from about three per cent to 22 per cent. It looks like sleep duration variability really isn't very heritable, but the number - how - number of sleep episodes which represents how much you move during sleep seems to be quite highly heritable. Why is this all important? Why are we interested in identifying the genetics of sleep patterns and sleep disorders? Well, at the most fundamental level it's trying to understand the molecular and chemical mechanisms behind sleep and trying to understand why some people need more sleep, and why some people prefer to sleep later than earlier.

It's well known that genes and genetic variants identified in GWAS make better therapeutic targets. It's entirely possible, with further work, that some of the genes that we've identified might make good targets for medication that might improve your sleep or might be used to treat sleep medically, sleep disorders et cetera. The thing we're most interested in, this ties back to our work with diabetes and obesity, is, if we have genetic variants and genes that are linked to our sleep measures, we have instruments for instrumental variables analysis, which allows us to start to pick apart the complex relationships between poor sleep and poor health outcomes. It's short but sweet; I'll stop it there.

I just want to say thank you to everyone who's been - who's contributed to this study, especially Vincent van Hees, who has really done a lot of work in improving the algorithms to get them to work with the UK Biobank data. Of course most importantly of all, thank you to the staff and participants of the UK Biobank for actually making any of this work possible.

**[END OF TRANSCRIPT]**