As Nick has alluded to, we've done a lot of work together as a working group carefully pre-processing all of the raw accelerometer information to get out some good stable measures, the physical activity intensity, and that has taken a lot of careful work. What I'm now going to describe over the next ten minutes is additional work that we're doing to try and further enhance the Biobank resource for you, and other researchers around the world who use this, that fabulous resource Nick has already introduced, and that it's worked on by other researches showing fabulous use of the research. Another example I quite like was done by Danny Smith and his group up in Glasgow using the drive measures we put back into the resource to try and get measures of [unclear word 0:00:47.7] rhythmicity and showing some nice, strong cross section of associations and mood disorders as well. This is a resource that has been used to try and answer questions for many different disease outcomes as well, and exposure types.

The work I'm going to talk about is work that we've done to try and get it more functional or behavioural measures from the accelerometer data such as trying to identify when and how often individuals might be sitting or walking, or driving or sleeping, and how we try and achieve this as via a statistical machine learning, essentially supervised machine learning exercise from the raw accelerometer data trace that you can see on your left. Is there any patterns or signatures inherent in this that can help us identify behaviour states of interest? They have been some past efforts with this, but there have been two fundamental flaws. Firstly, on the technical point of view the choice and models has often been inappropriate for the type of date, and, secondly, in terms of the training and validation of these machine learned algorithms, it's taken place in unnaturalistic laboratory environments. What we'd really like to know is what participants are doing in their everyday free-living lives.

To achieve that, in Oxford we asked 150 participants to wear a camera-based device as they went about their everyday activities, and we weren't so interested in trying to get a representative set of people, what's more important is getting a representative and diverse set of behaviours in which to train the model. The camera-based device looks a little bit like this that people wear via neck worn lanyard and the data then we get back looks, as you can see in the video here behind me, whereby from the camera data, so it takes a first-person point of view image every ten to 15 seconds, and we can see quite clearly whether a person is walking or whether they are sitting, or whether they're driving. We can even capture other lifestyle and health behaviours such as whether they're having a sneaky fry-up at particular parts of the day as well. Then our challenge, and interesting this from a Biobank perspective, is all these participants had the same accelerometers in the UK Biobank and we extract out a series of time and frequency domain features. These ones are orientation and invariant.

We run those through what we call a random [?torus 0:03:15.6] classifier to make some preliminary predictions on the type of activity the person is doing, whether it's sleeping, walking, running, and because for the time series nature of this data we can apply a second machine learned approach namely hidden mark-off

models then to smooth over those predictions to further improve the accuracy of the classifier we've done. This is work done in collaboration with Chris Holmes, who is one of the founding directors of the Alan Turin Institute.

What we had this in this particular paper published below is from 130 people, we had 160,000 minutes of annotated human behaviour and free-living scenarios. It's around 15 to 20 times larger than any laboratory type data set, and the key thing is the environment in which the data is collected, it's not a constrained laboratory environment. The overall accuracy then to identify these different types of activity then was 87 per cent, particularly getting good success in terms of distinguishing between sleep and sitting behaviours, and we can also identify activity such as walking quite well too. That's one form of validity around this machine learning approach we've taken.

We've also then applied this model to the Biobank participants, so 100,000 UK Biobank participants, and we wanted to, first of all, see whether there is any face validity to this, so does the output of this classifier make intuitive sense? We've got two different plots to show you here. On the left-hand side we've got accelerometer predicted time spent asleep, so x this is our day, y axis is probability of a population group being asleep, and we've got two different groups being shown here, those who self-report themselves as being definitely a morning person versus those who self-report themselves as being definitely an evening person, and we're getting very distinct differences, for example, at 8:00 am in the expected direction.

Then, over at the right-hand side, we've got what I like to call the Batman Plot, whereby we've got accelerometery predicted bicycling time whereby we've split up the groups by those who self-report themselves as being cycle commuters versus those who commute to work by other means, and, again, we're getting big spikes in cycling related behaviour at key commute times. This gives us further confidence that these machine learned models have got a face validity as well as the validity in our detailed ground truth data set in Oxford participants.

That has all helped us, as I said on the 100,000 UK Biobank participants. Once we do some QC this particular plot is 91,000 participants whereby we can build up this 24-hour model of human movement behaviours as identified via statistical machine learning. Of course, the beauty of UK Biobank is that all these participants have got a range of other variables as well, including the imputed genetic data, so working along with Cecilia Lindgren and her team at the Big Data Institute in Oxford too, we've been able to identify a variance then associated with, for example, sedentary behaviour status and also those are identified with sleep duration. As we would have seen just after the lunch in the afternoon, Sam Jones from Exeter and Mike Weedon and their team have done some really nice complementary work as well on trying to get out other measures of sleep related phenotypes. We're getting very similar stories in both forms of analysis that some of the sleep duration [?lowsi 0:07:14.0] overlap with our prior knowledge from self-reported G loss of sleep, but we're also identifying new lowsi as well, indeed, looking at the overall activity traits with the UK Biobank.

We've got here is a Miami plot, think of it as two Manhattan plots, so in the bottom we've got genome-wide association studying and self-reported physical activity status in 351,000 participants. On the top side

then we've got a genome-wide association study on the accelerometery derived overall activities rate, 91,000 participants. Unsurprisingly, we still identify more lowsi use in the self-reported metric because we had a much greater sample size. However, with the objective trait on the top, we're getting much higher estimates of heritability, and also very clear signals of chromosomes 10 and 17 that we hadn't seen before using the self-reported data. Again, I think this is a wonderful example of having subjective and objective measures at scale, and this, again, where UK Biobank is just truly unique in the world.

These lowsi were identified then allow us to understand more around the aetiology of movement-related behaviours, for example, [unclear word 0:08:38.5] heritability showing that the centre nervous system is playing an important role, and quite expected given its behavioural traits that we're dealing with, and further more those that were identified are proving to be I think better than we would have originally expected as instrument variables from [?indelium 0:08:57.6] randomisation analysis to try and unpick potential routes acrossality [sic]. For example, our preliminary MR is suggesting that an increase as an overall activity are associated with the reduction in diastolic blood pressure, and also hypertension, and conversely that increases in sleep duration are associated with increased odds of hypertension. A more careful follow-up will need to be done in this particularly in terms of replicating in other cohorts as well. Again, I think it opens up very exciting opportunities to further understanding of the cause and consequences of physical activity and sleep duration type behaviours.

To summarise, as Nick has alluded to, these metrics that we've been working together in a working group for the last few years are all available, and already used by researchers in the Biobank resource. The machine learned genome types and sleep duration, sedentary behaviour and walking that I've just presented, they will be returned back into the resource, so hopefully then further enhance it for other researchers to use for their particular research questions of interest. Of course, there is many exciting possibilities to further enhance the resource. For example, the Alan Turin Institution have given us some funding to look at unsupervised learning and to see other new health relevant patterns of activity that we might not have thought of before. With that I'd stop and thank you very much for your attention.

[END OF TRANSCRIPT]