The first one will be given by Nick Wareham from the MRC Epidemiology Unit in Cambridge and he will tell us about some of the kinds of insights emerging from use of accelerometery data at large scale.

I'm going to talk about the accelerometery and first to talk about physical activity and the summary variables that have been derived. The point of this is that we have self-reported physical activity assessed in Biobank, as in many other studies and that is what it says it is, so it is self-reported and in a sense that is correct. The problem comes when you try and derive measures of total physical activity from self-report and it's that transition that's very difficult to do. The objective assessment of physical activity on scale is an adjunct to the questionnaires, it's not a replacement to it, it gives you something additional. I want to show you what has been done in UK Biobank, because it is the biggest study of its type in the world, so 1003 people participated in this element of the project and the quality of the data is unbelievable. Often with accelerometery you have periods of missing data and that creates a real problem in this field because you can't distinguish between not wearing the monitor and being sedentary and a lot of studies have really fallen down because of that.

In the UK Biobank data the vast majority of people had data that was of sufficient length without this problem of wear time, that we could include everybody, we only had to exclude 7000 people and even that was actually a very conservative exclusion. This is extremely high quality data and 96,000 people is by far the biggest study of objective physical activity ever undertaken. Now we could have taken an approach of just letting that raw data on accelerometery go into the showcase and allow everybody to do whatever they wished to do with it, and indeed that has actually happened. I want to argue that there is, because this is novel type of data, I think it's important that we get a consensus view on some key summary variables that can be derived from that data and get them back quickly into the showcase for use. To avoid problems of interpretation, which I will show you something about a little bit later, but also people want to use this data either as a co-variant or as an outcome or as an exposure, and we need to get the variables back into the data very quickly.

We did that as a consensus group and all the processing alga rhythms are available online, and essentially there is some pre-processing of the data to put it in a usable form, it's measured 100hertz in three axis, this is triaxial waveform accelerometery on the wrist. Then there's a series of post-processing and feature extraction and prediction, this is where you get into inference. Again, all of this, all these steps are open-source and documented. In the paper that [?Aiden Leaden 0:03:48.8], who is going to speak a little bit further about this in a second, we were able to give this summary variable, which are called average vector magnitude. In essence this is a summary variable across those seven days of total physical activity and you can see, as one might expect, it declines with age, so this is a useful variable for the assessment of total physical activity. There's an interest in physical activity in terms of time spent in different intensities, and this field has been bedevilled by people having different cut points and different definitions of what they mean by light or moderately vigorous or vigorous physical activity, and that's really deeply unhelpful.

We decided to not get into that debate and come up with some artificial thresholds, but rather just release the data and give the distribution by individual of the intensity, and in essence allow the data to tell us what those cut points should be, because in time we'll be able to relate time spent in this type of activity, sedentary or other types with outcomes. That's what should determine these thresholds, not some arbitrary decision, so let the data speak. These data are now back in the showcase, the source data, right back to the original raw accelerometery, through to these derived variables about acceleration and intensity. It's really encouraging that they have been used in exactly the way that we predicted, not just as co-variants but as outcomes, so people who are going [unclear word 0:05:32.1] of physical activity. Also as exposures here looking at the relationship, doing one study between objectively measuring physical activity and brain structure, combining both the activity and the imaging. Looking at different intensity cut points and how they relate to bone health, to link back to the previous talk. Even in this study published in circulation earlier this year, at the bottom here relating objectively measured physical activity to mortality.

Now the numbers are small because the people who had physical activity's only one fifth of the total UK Biobank and that measurement was done towards the end of the projects, and the follow-up time is small and you can debate whether this is likely to be reverse causality or not. In the future this is going to be a very important feature of the UK Biobank as further events emerge. Now what is important is that we're also doing a repeated measures sub-study, so that this will allow us to estimate the regression dilution factor and therefore correct observed associations for measurement error, because physical activity will vary between periods and we only measured it for seven days. Therefore we're doing seven day accelerometery repeated four times across the year at three-monthly intervals. Three thousand participants have signed up to do that and the very encouraging thing is that we're now in the second phase of this, this is a real example of what's brilliant about UK Biobank, is the team manage this process absolutely splendidly and it's very, very efficient.

A high proportion, I think well in excess of 95 per cent of people are participating in every wave, so this is going to be a really interesting dataset, to calculate regression dilution factors but also to look at seasonal variation in physical activity, which we don't know too much about. We also need to develop further consensus work, so Aiden's going to talk about some machine learning approaches that try and characterise types of physical activity. One of the things you need for some of those machine learning is to know how the accelerometery is orientated on the wrist and indeed which wrist it is on. Those things could be important in determining various parameters, they don't matter for effect of magnitude but they might matter for some of your feature extraction. The important thing in UK Biobank, people were instructed to wear it on their dominant risk, they didn't always, and there were two types of packaging of the monitor which allowed for some of the monitors to be orientated at different angles, and some of them indeed to be inserted upside down or back to front. There needs to be a process of inference to try and work out the orientation of the monitoring space if you're trying to make inferences about what activities people are doing.

The other feature that can be extracted obviously is sleep, because people wore these monitors for 24 hours and it's possible obviously to see inactivity particularly at night. There are a number of alga rhythms that

have already been published by different groups using the UK Biobank data. Rather than come up with a consensus sole alga rhythm as a first effort, what the sleep working group is trying to do is focus on comparison of different inferential alga rhythms with criteria methods, with more gold standard. Not to say there's only one way of doing this but just to compare how people have done them. I just want to finish by saying something about other related traits, which I think exemplifies why having this sort of working group consensus development is important. Now I want to talk about cardiorespiratory fitness, so if I go back to that paper I mentioned which showed the relationship between physical activity and mortality, it also has cardiorespiratory fitness, which shows a very strong inverse relationship with physical activity. In that paper it was calculated as [?V02 max 0:10:04.3], net oxygen consumption as a function of workload, maximum workload and weight.

They got the workload estimates from the showcase, but in the Biobank we actually did an individualised risk stratified ramped protocol, and that's because we wanted to be careful, this is a general population, we want to be careful not to cause any adverse events. We actually changed the protocol that people were going to have according to risk factors, and actually if you then look at the variable that's called maximum workload, it has this distribution. Now if you look at that, that does not look physiological, okay? There are actually only 22 different points that somebody could have, different workloads and they are the predicted workload, the maximum that would be allowed by the protocol before it was stopped. That is largely driven by the Rose Angina risk equation, so this isn't fitness at all, it's cardiovascular risk being used to predict whether somebody should have an intense or a slightly more lenient fitness test. It also only, we only took people up to about 50 per cent of max and you can see there, if you compare it to what it should look like, which is on the right, it should be a normal distribution if it were truly max workload, going up to about 300.

I think this indicates a real problem which is you kind of, it's not the individual research's fault because they took a variable, they thought it was something. It's not the journal's fault because it's virtually impossible for them to pick it up. It's not Biobank's fault, it's probably for us as a community to try and come up with some collective ways of trying to help people avoid misinterpretations like this. I think that probably goes for a multitude of areas and maybe it's something, [?Rory 0:12:12.8], we can come back to in discussion. I'd like to think that this accelerometery working group has really helped the field, that we've pushed things forward and it's very encouraging to see the amount of work that is now being published on this enormous resource that's being collected so well. I'm going to stop at that point.

[END OF TRANSCRIPT]