So could I invite Nick to lead us? You sit in the middle. You're not to be drowned out by all these rather grand people. Could I invite Andrew, Cathy, Rory, and Naomi to take a seat for the question and answer session? I'm assured that there are some microphones over there, although clearly my eyesight is deteriorating to the extent that I can't see them, but Rory has one, brilliant. Okay, I can't see all that brilliantly well because the lights are shining in my eyes, but are there any questions from any of the speakers this morning? Microphones will be provided. There's one, number two - if you could just say who you are, that would be great actually.

M: I'm Anthony Khawaja. I'm an ophthalmologist working in London. You asked about ways that maybe we could improve the resource. I know this is something some other studies have done around the world is the opportunity to enrich the number of cases in your resource, so I study glaucoma. There's a certain amount of glaucoma in UK Biobank. It's probably relatively underpowered to discover the things we want to. Can we enrich it? Can we add more glaucoma cases to the resource? What we'd need for that if we fund it ourselves if we discovered glaucoma cases in our clinics and can we put them through the same process that the UK Biobank people have gone through? Can they have the same email questionnaires, the 24-dietary recall? Can they have the same genotyping done at the same lab, the blood tests and everything so that later on we can put these together and have more power to discover things for whichever slightly rarer disease you're interested in?

Thank you, maybe Rory could take that.

Rory: So UK Biobank itself is a cohort of a half a million people and so that is, if you like UK Biobank and it's not intended, although it would be very desirable for it to be bigger, for us to recruit more individuals into UK Biobank. I'm often asked why UK Biobank is half a million people to which the answer is there wasn't enough funding for a million. So the direct answer to your question is we're not going to be recruiting more into UK Biobank. However, what we are trying to do is provide tools that help others to do the kinds of measures that have been done in UK Biobank. So the vision measures in UK Biobank from the baseline are a really good example of what I was talking about. We started recruitment. There were no eye measures and a group from Moorfields came along and said, 'Why are there no eye measures?' We said, 'Well what do you want?' They said, 'We'd like to have visual acuity.' We said, 'Well, yes, but be ambitious. What else would you like?' So they said, 'We'd like to have imaging.' We said, 'Okay, what else would you like?'

So they said, 'Well optical coherence tomography.' We said, 'Okay, fine you can have whatever you want in ten minutes.' So they went away, they pulled together a protocol that could be done in ten minutes, and the funders didn't believe it could be done, until they actually pulled people off the corridors and videoed it and we sent in the video and it was funded. So we have optical coherence tomography on about 100,000 people. I think that exemplar is then something that can be used by other big cohorts that are being

established, like the All of Us cohort of a million people in the US that is looking to UK Biobank to the things that worked to increase the numbers there, so that we can get larger numbers in these sharable cohorts. Similarly, with respect to some of the web questionnaires, having developed them in collaboration with various scientists, what we're now doing is trying to make those web questionnaires a generic tool. So Alan Young and the IT team are trying to take them, so they then could be taken by other researchers and used for those studies. So we're trying to generalise in that way.

Thanks Rory. Do any other panellists want to contribute to that? Thank you. There's another question down here.

John Todd from Oxford. Is it timely to approach the funding agencies or whoever to ask the children of current participants to participate in a further study?

Good question. That might be one for Naomi.

Naomi: Well, I think it's really up to the funders and I think if there was an appetite to recruit the next generation into a study that looked very much like you could biobank - as Rory said, recruitment into UK Biobank is complete. It would involve setting up a new study. If the research team you see would think that's of value, then that's something we should discuss.

Rory: This is something that's being considered by All of Us the US cohort. Initially they were saying, 'Well, we're going to recruit everybody.' I said, 'First of all, make sure you can recruit anybody and focus on the adults, but then that would provide an opportunity to recruit through the parents a younger generation.' So I think it's an attractive model and it may well link in with plans that again I think John Bell might have mentioned, which is the possibility of having a cancer detection cohort in the UK, where individuals have a blood sample collected and then every few years they have a blood sample collected to try to detect cancer through circulating DNA. There are discussions around where UK Biobank could be part of that, but also the organisation that Tim Peakman, our chief operating officer previously set up at biocentre, which was established to help support these kinds of projects. Biocentre would be setting up these recruitment centres, and if those recruitment centres were recruiting or reassessing UK Biobank participants, then there is certainly the potential that they could recruit into new cohorts using what we've learnt from UK Biobank. Sorry Cathy.

Cathy: Yes, it was just to add that we have had that question to us before from a number of scientific groups and it's one that we've discussed a bit and it's an issue of not just funding, but bandwidth and prioritisation and so on. So once there's a sufficient scientific rationale to take something forward and a following wind and a coalition of funders and so on brought together, then these turn from a wonderful idea into a reality. It's one

that might in due course and it's just a question of when the time is right I think.

John: I think the scientific rationale is unassailable, not least the recent paper in science from [unclear words 0:08:18] illustrating and proving how non-transmitted [unclear words 0:08:24] have such a large impact on the phenotype and behaviour of the offspring, and also explaining a large proportion of the familial clustering of common traits and diseases.

Thank you. Are there any other questions? Do we have any of the funders here who might want to comment on the way forward in terms of these new initiatives? I hate to put the MRC and the Wellcome Trust on the spot like that, ho, ho, ho. I think whilst you are coming up with the next question - and there's one over there - a comment from me, what might not be widely appreciated is just how good UK Biobank became at measuring routine simple things. You may assume that measuring the serum marmalade level is a straightforward activity. What UK Biobank to its eternal credit realised is that most things that you measure prove to be a lot more difficult to measure reliably than you ever anticipated. I have a little sneaky sadness, I must admit that we stood down the team who were in Cheadle who did all the routine blood testing and metabolomics if you will. It would take a bit of effort to reboot that activity. We've stood the team down and that's a shame that that had to happen, but of course the knowledge is there and that can be restarted. There's a question over there sir.

Yes, this is Johan [unclear word 0:10:18] from the European Bioinformatics Institute. It's really to Catherine and Andrew. Cathy, how thought through is the exposure of the routine healthcare record data in your own head? So there's one which is the very raw data and there's the other extreme which is a kind of consensus expert group this person definitely has dementia or something like that by fusion of that. Is there an intermediate space between those two extremes, because I think many people like myself don't want to try to understand the raw data, but want slightly more dimensionality than the final pull? After you've answered that, maybe I can ask Andrew whether that gets scaled out across all the UK or not.

Cathy: Yes, so you know there are a number of degrees in between those two extremes, so actually the far extreme of having every case expert adjudicated I think is a nirvana that we are unlikely to reach for many years to come. So we shouldn't really strive to try and do that yet. So we've only done that in very tiny relatively speaking subsets in a way that we think will be reasonably generalisable to assay the accuracy of the much more massive data set. So the construction of the algorithms is really just a grunt work task of pulling together code lists using explicit methodology and it's really pretty easy for ICD which has got a few thousand codes in it. It's a real bind for read code, which is what's used in primary care because it's got over 300,000 codes in the latest version, which includes signs and symptoms and similar concept coded in multiple different ways, which is useful in practice, but really difficult to use in research. It's just a grunt work task really and so

the more one can crowd source that but using it as common methodology, the better.

In the interim there are all sorts of things that can be and are being done, just by using codes and dates. What we're trying to do in the interim is to make the first code of any particular type that occurs in its associated date, available in a more digestible format than is currently the case, which allows the big studies that GSK for example are doing, to look at the associations of genetic variance across a wide range of different disease phenotypes to help them in their drug discovery efforts much faster. They've already been cracking on with that, but we're just trying to make it easier by degree really. Does that answer your question?

Yes.

Andrew.

Andrew: So it's a good question and we shouldn't underestimate how challenging these data are. They're very messy, huge heterogeneity, often lack of standardisation. A lot of folk didn't believe that. A good example is Google DeepMind working with Moorfields. They said it took them 14 months to sort the data out and a day to run the algorithm. So I think we've just got to put a sense of realism round that. Saying that, with these new tools, you can start to try and forge convergence and standardisation because that's what we need. I would argue the UK needs a clearing house of these tools. Rory, Cathy and Naomi ran a great workshop two or three years ago where we had Nigam Shah over from Stanford. They've got a data science services team that calculate the positive predictive value of specific phenotypes, so that researchers know the validity of the algorithm. I think that's what we need to do, but start with a few selected datasets, prove the model and then scale.

I guess history teaches us that with any datasets, when it becomes realised that people are actually going to look at it and use it, the people who are generating it in the first place tend to pay a little bit more attention to the quality that they are putting in in the first place. So I think that actually UK Biobank will drive all sorts of improvements that we at the moment probably don't quite anticipate fully. Are there any other questions? There is number three over there I see at the back.

F: Hi, Angela Hodges from King's College London. We've talked about health risk mainly, but I was just wondering if you could comment on the depth and quality of data for protection, because by definition it will be sparser because you're not expecting to see GP and that type of health record for these people.

Could you repeat that?

F: I'm interested in comments on the data, the quality and the depth of the data for actually looking at health

protection for those people that are less prone to these diseases that we've been talking about, because that's an important alternative side of the coin.

So health protection, Cathy anything.

Cathy: Yes, so talking off the top of my head I guess you mean to monitor the other extreme, the very healthy people regarding a range of different diseases or phenotypes. So I guess the main way in which we are really going to be able to achieve that, other than by the absence of appearance in health records – and one might do it by looking at the absence of multiple different morbidities for example. So there are ways you could do it through the health records, but perhaps also through the web questionnaires where we'll be looking at measures of function, quality of life. There are some baseline measures of course that are highly relevant to this that help us to look at, for example, resilience against mental health disorders, happiness, self-rated health status, which does seem to correlate very well with other measures of health status that are apparently more objective. So Naomi may like to comment more.

Naomi: Yes, I think that's right Cathy and in terms of the web-based questionnaires, we will be focussing in the next year or so on things like quality of life, well-being, as you retire, happiness and those more subjective measures on how people are feeling generally, both their physical and mental health. I think that's probably the best way to capture well-being rather than through coded data through the linked records.

Thank you. There was another question, James Peach.

James: Thank you very much, James Peach from the Medicines Discovery Catapult. I noted in the talks that there is a theme about trust essentially and maintaining the trust of the participants. There has been increasingly, announcements about commercial use of the data, so the Regeneron deal and Perspectum Diagnostics and others. I'd be interested in comments from the panel about how you managed both trust and commercial interest given the concerns from the public around commercial and how you see this evolving over time.

Rory, would you…?

Rory: At the beginning of the study we made it very clear to participants that the resource would be available to academic and commercial researchers on the same basis and the consent was very explicit about that, but of course they were recruited in 2006 and 2010 and people forget what they sign. So Andrew Trehearne, our head of communications, I think had done a fantastic job in ensuring that participants knew what was happening and why it was happening. So in particular with the arrangement with Regeneron to do the Exome

sequencing, a lot of thought went into how that was presented and explained to participants. We had great support also from the funders, from the Wellcome Trust and the MRC, again around how one can explain the rationale for this. Andrew has got a microphone, so he can comment, but the reaction was very positive and when we go and talk at meetings with participants, again in the same way they respond as Alex has mentioned, 'Well, of course you should be having access to my data.' They go, 'Yes, I understand that. That's the way to develop new drugs, to improve health.' So I actually think the participants understand it if it's explained what the rationale is and how it would benefit the resource as a whole. I don't know Andrew if you wanted to expand on that.

Andrew: Just to underline what you said really, we hold very regular participant meetings. Two hundred or three hundred participants come out for afternoon tea or a glass of wine and a sandwich after work and we talk about these issues and we discuss them with them. People are very supportive. We know they're a special bunch of people. They went out of their way ten years ago to join the project, which was two or three hours of their life. Many of them have come back to be imaged and taken part in the questionnaires. So at every occasion we do our best to remind them what it is that they're doing, why it's important and I think generally speaking people just want us to get on with it.

Rory: Participants can withdraw at any time from the resource. I can't remember what the exact number is now, but of the half million it's probably of the order of a thousand or so, something like that. So again I think we're getting the communication across. We can always do better and work out ways of telling people. Of course now we've got some great stories. We've got the story that Nick described of how the resource is being used and that is exciting participants and I think encouraging them to see the value of being involved. In fact, we get quite a lot of people saying, 'I wish I could have joined' or 'I wish I had said yes.'

Andrew…

Andrew: So it's a good question James and also a key issue for the life sciences industrial strategy which is predicated on the NHS really driving innovation. So I think there are a few principles, here, firstly of transparency, so that's the first thing, the rule of no surprises, but also the concept of benefit sharing. I think if you look at the social science in this space, the notion of benefit share being explained, and the public interest test being asked, because benefit sharing can be monetary, but it can also be non-monetary. I think what the public is looking for is a fair explanation of the benefits. How does UK Biobank or the NHS or society benefit from such exploitation?

Thank you, Andrew. Now, I'm going to draw the session to a close there. I hope you've all enjoyed it enormously, but before I ask you to thank our speakers again, I've got a request. So I would like all the people

who work behind the scenes for UK Biobank in Edinburgh, in Stockport, in Oxford and in Cardiff to stand up. There must be some of you here. Now I would like to say that 12 years into this programme these people have done this country a massive service and I think as well as applauding our speakers, they deserve our enormous respect and thanks. So well done you lot.

[Applause]

[END OF TRANSCRIPT]