[FEMALE RESPONDENT]
[Other comments:]

Thanks very much, Alex. I hope that little slip of the tongue, where you referred to me as Sally, wasn't some kind of weird premonition. I want to tell you a bit about following the health of the half a million participants, because a big part of our job has been to really try and make this a prospective resource. To make it a prospective resource involves linking all of the very broad range of data that you've started to hear about, that you're already working with, and that you'll hear more about, through to prospectively ascertained health status. How do we do that for half a million people? It certainly needs scale. Our aim is, of course, to identify a broad range of incident disease, and other health-related outcomes, to subserve the needs of a very broad range of research studies. There are traditionally a number of different approaches to this, but many cohorts have used active methods, requiring participant reengagement, to find out about what's going on, prospectively, during the course of follow-up.

That has previously been done largely using face-to-face reassessments. More recently using, first postal, and then web-based surveys and certainly, in our experience, using web-based surveys is a very rapid way of acquiring data and then providing it directly on to researchers. Many of you have been able to take advantage of that. It is, of course, prone to incomplete coverage, and selective loss to follow-up, perhaps sometimes for the very interesting participants who you want to find more about. That might be particularly the case for cognitive function, where those who have the least cognitive function, and are the most likely to develop dementia, will be the least likely to engage with these types of methods. Nonetheless, it's an extremely useful way of gathering types of data that you can't get from other sources. I mean, another issue is that you may miss cases of disease emerging between assessments and, as memory declines with time, you have to do these things frequently, or very carefully, to be able to maximise the use of the information.

Another way of gathering information is using passive methods, via linkages to national health records, and that's what I'll focus on mainly today. That means we can follow all participants, at least in the UK, without the need to reengage them. We like to reengage them, but we don't like to have to reengage them. It should be efficient and cost effective. It is relatively efficient and cost effective, actually, compared with a lot of other data gathering exercises, but it's not nearly as efficient and cost effective as it should be. It does require, or certainly benefit, from broad consent at recruitment, which we have in UK Biobank. It relies on a universal healthcare system and we should not underestimate the advantage we have in the UK in still having that, and it requires the availability of relevant datasets, preferably at national scale, because that makes life very much easier. We can only detect cases of disease that are diagnosed in a healthcare setting, using such methods, so that is an important limitation.

The data need to be accurate and sufficiently detailed for research studies and, at the very least, we need to know how accurate they are, and how detailed they are, so that we know for what they can be used. Just briefly to let you know a bit about the web questionnaires that we use during follow-up. We have email addresses on about 350,000 of our participants and we use this as a method of re-contacting them to gain information on, firstly, more detailed assessment of exposures, and increasingly to obtain information on

outcomes that we can't ascertain through linkages to healthcare records. Each time we do one of these exercises, and this is work that's coordinated largely by Naomi Allen in Oxford, we get responses back from about 30 per cent to 50 per cent of individuals, and that does appear to be increasing over time. Perhaps due to participant networking and knowledge of the fact that these questionnaires are out there. Perhaps due to the use of reminders, and perhaps due to the use of better, more excitingly designed questionnaires that are more engaging for participants, so we've learned from all of these things.

These are the areas that we've been looking at with questionnaires, so we've done some early questionnaires on 24-hour dietary recall, which went very well and really paved the way for subsequent questionnaires on cognitive function, which will be repeated to track change over time. Occupational history, mental health, and digestive health, and you can see the increase in completion rates, from amongst those administered a questionnaire, over time. We are also planning a further round of questionnaires over the next few years, ascertaining pain, quality of life, functional status, and some other further exposure information, as well. In terms of following the health through record linkages, which is the passive follow-up method that we focus on, we are now in a position, after quite a few years of wading our way through the regulatory and bureaucratic quagmire that exists around the UK, to obtain regularly updated information on a wide range of diseases.

We focused our efforts on datasets that will yield maximum information across the widest range of diseases, not yet on disease-specific audits and registries, although those are very much on our radar for more detailed phenotyping. Mortality data comes in the form of death registrations, which is a mature system in all three countries where we have participants, and you can see we get large numbers of data very rapidly, and we update these about annually, or so. Similarly, we have a mature cancer registration system around the country and, through the work of the National Cancer Intelligence Network and others, including academic groups, these data are becoming more and more detailed, and yield more and more information about stage, grade of cancer, and so on, as well as the date and the diagnosis. Then we get information from hospital records, as well, through national systems. The outpatient episodes are not generally diagnostically coded.

It's the inpatient episodes that yield the most information, so it's really about hospital admissions, here, and that gives us information on thousands of cases of a whole range of different diseases. Mainly incident diseases, but also some retrospective information about prevalent diseases before the time of recruitment. Then we have spent many years, as Alex alluded to, battling to try and obtain linkages to primary care data, because there has not been a single national resource to obtain this information from, in any of the countries. The most rapidly growing comprehensive resource is in Wales, which now has coverage of 80 per cent to 90 per cent of practices for all types of research and has a particular focus on obtaining information for us for UK Biobank, but you can see that that's a relatively small proportion of our population. We've worked through system suppliers and various other software companies to get hold of the data that we have. We're absolutely intent on making sure that that coverage grows rapidly to much nearer 100 per cent, and I'm very happy to discuss how that works with any of you, offline, if you're interested.

These are messy, real world data and they're not collected primarily for research and we know, just from being clinicians, or having worked with these data before, they're not 100 per cent accurate. They're prone to some administrative errors, which by and large are rather small, and to some clinical errors, which tend to be a little larger. They're mainly structured coded datasets. They use the International Classification of Diseases system and, in primary care data, a much more complicated coding system that's currently called READ and will map over into SNOMED in the near future. We've worked with these data as we've pulled them in to try and make them as useful as we can for researchers, to try and make them available in as useful a format as possible, but we've also tried to make them available as quickly as possible. So those of you who have grappled with these data, will have had to work hard with the early releases, because we just made them available in the very messy, almost [?raw/RAW 0:08:30.0] format that they came in, having performed some basic checks.

We're trying to make them much more useful to form, to create disease flags that combine the data in algorithmic ways that are intelligent, and sensible, and maximise accuracy for a range of research purposes. We will continue to provide them in messy or raw format, as well, so that those of you who want to go it alone and share your results back with us, can do so. Hopefully, that way, we'll grow the usefulness of the resource quicker. It's important for us also to consider the limitations and potential additional sources of data that we might pull in, to enhance these structured datasets. In terms of cancer, these are some of the numbers of the commoner cancers that were recorded in the cancer registries by the time of recruitment, or that we predict will occur during the course of follow-up, or indeed have observed during early follow-up. Going beyond the information in the registries, though, beyond that structured data, we've been working on the feasibility of obtaining additional information for sub-typing cancer cases through regional linkages, to such information as free-text histopathology reports, digitised histopathology slides and actually obtaining tumour specimens, which will be very interesting and helpful for a wide variety of cancer studies.

We've been working, in particular, with the Universities of Newcastle and Leeds, so far, in this regard, but we hope to roll that out more widely as we learn how to do it. We've also been thinking about how to phenotype a wide range of non-cancer diseases and one of the important things here is the role of primary care data, purely in ascertaining the vast majority of the cases that are out there. Here you can see the numbers of prevalent diseases, new cases observed by 2016, and then cases predicted to occur over through to 2026, from the sources of data that we have across the entire cohort. Self-reported recruitment, hospital admissions, mainly prospective, and death registries, again, prospective ascertainment. Pretty chunky numbers, but we know that these will not capture all the cases. If we take the data that we have from primary care, which is for about half the cohort, and use that to extrapolate to what we would expect to see across the entire cohort, you can see a substantial increase in the numbers, and that is particularly the case for conditions such as dementia. For example, where the numbers more than double.

For conditions like chronic obstructive pulmonary disease and Parkinson's disease, similarly. The numbers for stroke increased by about another 50 per cent and the numbers for myocardial infarction,

somewhat unsurprisingly, don't increase a great deal, because we know most people with an MI will be admitted to hospital. This is really important and helps us to focus our efforts, and this is just a handful of the conditions that we're keen on looking at. Understanding how accurate and how complete these records are, is really important. Just to illustrate that briefly for dementia, we can do some work based on published studies, and understand that positive predictive values, to a large extent in published data, suggest that these routine sources give you decent accuracy. Not perfect, but decent, and we can understand a bit about different subtypes, as well. This is true of other diseases, too, so the positive predictive values are generally higher for Alzheimer's disease, for example, than they are for vascular dementia, which is, in any case, much harder to define and there's much less agreement about how to do so.

We've done some direct validation work, as well, within Biobank and this is just one small example showing that the information that we're obtaining from direct validation in subsets of the cohort, largely reflects what we've found in the literature. So reasonable for dementia, overall. Good for Alzheimer's disease, not so good for vascular dementia. We're doing similar studies across a range of different disorders, and then providing that information to researchers to they can use the information that we give them in the context of its accuracy. Just to give you a flavour of what's coming and how we're making this available to researchers, here I've shown several conditions that we're looking at, and trying to create these disease status, algorithmic ready-to-use data fields, within the dataset, combining across the various sources of linked data. In green, I've shown the conditions for which we've got those disease algorithms already up and running and available for researchers, and they're being used already.

In orange, I've shown the ones where we're just about ready to go, our algorithms are advanced in preparation, or ready to be released in the immediate future. In red, I've shown the disorders that we're working on, but it's going to take us a little while before the algorithms are complete, and the time that that will take varies. By the end of 2018 we think that things will look much more like this. They might look a bit better, but they should look at least as good as this, where we'll be releasing, at least for the datasets that don't include GP data, a lot more off-the-shelf algorithms. Then, by the end of 2019, it should look a lot more like this, where we'll still be working on some conditions, but we'll have plugged most of the gaps for many that we've been working on for a while. We would like to move faster and we would like to appeal to the community to help us in that work, so that we can really increase this and make it a very big green and complex slide that covers loads and loads of conditions.

What about the research uses of the linked healthcare date? You've seen Naomi's slide and I think that's a very useful summary of how the data are being used. We've seen a wide range of study types, mainly cross-sectional associations to start off with, not really using the prospective data, but it wasn't available. Large-scale genetic analyses have been common, and they have started now to use the healthcare data that we link to. The use of the healthcare data in prospective epidemiological analyses, as you saw from Naomi's work, has focused mainly on mortality, at least in terms of the highly cited publications, but increasingly we're seeing people now use cancer data and the hospital admissions data. As we pull in the primary care data, we'll

no doubt see more use of all of these. The extent of the use, I think, but this is anecdotal and we haven't done a formal analysis of this, depends on how long the data had been available to researchers, their pre-existing familiarity with the dataset and coding systems, the complexity of the dataset that we provide, and the numbers of disease cases and length of follow-up.

Improving all of those will increase the speed and usefulness of the resource. We've seen that the use of prospectively ascertained deaths, cancers, and hospital admissions is definitely increasing, and this is particularly the case where we've provided off-the-shelf outcomes, so it certainly makes sense for us to continue to do that. Going beyond the linked coded data, from structured coded data to unstructured data may well provide additional opportunities for phenotyping, as I've illustrated for cancer. It might require many unstructured data sources, and it's very difficult to know how to prioritise these, at the moment, and we certainly have to do a lot of pilot work in this regard. We could access, for example, regionally free-text of electronic records, but we can't do that nationally. Ditto for complex electrical signalling data, histopathology slide sets, I've alluded to, and clinical imaging data. Obtaining these at national scale is certainly challenging, and then working out what the incremental value is, is really important to know what the trade-off is between depth and breadth, in this regard.

It does need scalable approaches, crowdsourcing, perhaps the use of tools such as natural language processing, machine learning, and artificial intelligence, to make these data as usable as possible. If you would like to help, or make suggestions, this is your opportunity, please get in touch with us. Around data linkages and algorithms, please contact me, and if you have ideas or thoughts about web questionnaires, please contact Naomi. Thanks very much.

[END OF TRANSCRIPT]