F: Thank you very much Alex. So, all of this linked healthcare data and how do we convert it into useful information, particularly for research users who want to study large numbers of phenotypes but perhaps aren't so familiar with the healthcare systems that are used in the UK and with the challenges that those present. So you've seen slides similar to this before showing the four main sources of data that we have and the fact that we have coverage for death, cancer and hospital admissions data for our entire cohort and currently just under half the cohort for primary care information. The types of data that we have, you can see here summarised on this slide and they're relatively straightforward for the top three, so dates and information on cause of death, site, stage and grade of cancer or in the case of hospital admissions diagnoses and procedures. In the case of primary care information the coding systems cover a lot more types of episodes. So not only are diagnoses coded up, as are procedures, but symptoms and signs are also coded, so for example wheeze or low mood, rather than necessarily a diagnosis of asthma or a diagnosis of depression.

Specialist referrals are coded up as well as are prescriptions and laboratory tests in many cases. So that creates great opportunities and great richness, but also additional challenges. Many of you will be familiar with the coding systems that are used for the top three of these, principally, the international classification of diseases and most of the information that we have from these sources is in ICD version 10, which has been being used in the UK for quite a number of years with procedures in the hospital admissions dataset, coded to something called the OPCS4. In the case of primary care, the coding systems are much more challenging. So in the international classification of diseases, if you go down to the most detailed level there's about 10,000 separate codes. If you rack up back one level there's about a thousand different coded entities at a slightly lesser level of detail. That's reasonably straightforward to handle, as is the hierarchical nature of the chapter base coding system.

In primary care, in Read version 2 there are about 70,000 different codes for different entities across those different types that I've described. In Read version 3, which is now being used widely and is used in the largest amount of primary care data that we currently hold, there are 250,000 different codes potentially available to GPs to apply. Some of them look as if they mean almost the same thing, but they exist and it's important to understand that. So that adds a layer of complexity to how we need to help researchers understand these data. Then BNF and DM&D codes are different systems that are used to encode drug prescriptions. So how do we make these data valuable? So in the case of cancers, the mature cancer registration system in the UK is very valuable and it already provides high quality disease status information. So with the exception of looking at the added value of those regional resources that Robin was talking about, there isn't a great deal we need to do to make that easier for researchers to navigate and it uses ICD, which is generally well understood.

For non-cancer conditions, as Mark was alluding to, we plan to rapidly create disease status flags for around a thousand conditions based on the three digit level of the ICD-10 coding system. So that covers a very wide spectrum of disease conditions and using resources available through NHS Digital here in the UK, we've been able to use the mappings that are available within their system between ICD-10 as a well understood
spine and Read 2 and Read 3, those rather more complex systems that are available in primary care. We've also mapped information from participants self-report at baseline across to ICD-10 format and that's what's enabling us to do that. So that means we can combine information from self-report at baseline, the hospital admissions, primary care and the death registry information and map it all to an ICD-10 format. So that's not going to be perfect, but it will give coverage for a wide range of diseases.

So what are the limitations of these linked healthcare data for assessing participant outcomes? Well, first of all they're pretty messy real world data. They're collected at the clinical coalface. They are encoded by coders who usually sit in the windowless basements of hospitals and aren't paid very much money and that's where they come from. It's actually amazing that they are as good quality as they are and although they're not 100 per cent accurate, the rate of administrative errors is really very low. They of course are subject to clinical errors as well, and as you know, those of you who are clinicians particularly, patients may be initially diagnosed and labelled with one condition and then over the course of time that changes, but the code generally will stay in their records. So there are also conflicts within the record of any particular participant. So we will be using and deploying these comprehensive NHS Digital map-based tools to combine data across the different coding systems.

As I've suggested they're not perfect and they haven't been extensively validated condition by condition, but they will provide a very useful scaffold for researchers who want to simultaneously study many conditions or want to focus on one condition but take account of others. They will miss cases and that's because the mappings through to the primary care data focus on diagnosis codes. As I mentioned before, there are many codes that are used in primary care settings that may signify a diagnosis without a diagnosis label being applied. So, for example, the combination of wheeze and prescription of an inhaler might signify asthma or COPD or repeated referrals to a COPD specialist nurse assessment clinic might signify a diagnosis of COPD without that diagnostic label actually appearing in the record. For some diseases, quantitatively that can be very important.

Some conditions are not well captured at all by these health records systems as Robin alluded to and those include for example many mental health conditions or decline in cognition, prior to an established diagnosis of dementia and they do lack information on sub-phenotypes for many conditions. So it's common for either for diagnostic labels that specify subtypes to not be used and for a less specific label to be applied or for them not really to exist in a helpful way that is well understood by the biomedical research community or indeed by clinicians themselves. So to deal with some of those challenges, UK Biobank has been working with experts to create a set of more accurate disease status indicators that are algorithmic combinations of more refined sets of codes for a range of different conditions, to estimate the accuracy of these, to consider which additional linkages will add most value. That's an ongoing exercise which Robin alluded to. To think about what are the most appropriate scalable approaches for disease sub-classification.

So here are the areas that we've currently been working on for some time to create these disease status indicators from linked data to add value to the scaffold of the thousand or so phenotypes that are referred to.
So you can see that we've been choosing the areas based on what we think is likely to be of most interest for researchers and the commonest disorders that are likely to be of most interest from a public health interest as well. So cardiometabolic disorders/diabetes, MI cardiac disorders, renal disorders. You can see on the top right there, ocular disorders and then respiratory conditions, COPD, asthma, inflammatory, venous thromboembolism and we have a very active mental health group helping us to deal with that side of things. Musculoskeletal conditions including fractures, which Nick Harvey, who was at the back has been helping us with and arthritis and then dementias and other neurodegenerative conditions.

So just to give you an example, for example if we take chronic obstructive pulmonary disease, where do the cases come from? So in those participants from whom we have information from self-report at baseline but also through the linkages, from hospital primary care and death data, the algorithm that we've constructed contains about 15 ICD codes, more than 40 Read version 2 codes and more than 150 Read version 3 codes to take the landscape of all the coded data into account. You can see that if we look at cases that were detected prior to recruitment you can see that quite a large proportion of them are only detected through baseline self-report. If you look at incident cases that are detected after recruitment, you can see how important primary care data is as a source of this condition, with nearly 60 per cent of the cases being detected through linkages to primary care data alone, because they're not associated with a hospital admission or haven't been yet.

In terms of accuracy we've been reviewing the published literature to see how accurate these coded data are for identifying cases of disease and here's a summary of some of the work on dementia summarising what's out there in the literature, which in summary shows wide variation in different studies, but suggests the positive predicted value - that is the number of cases identified by codes that are true positive cases - is generally pretty high. It's pretty good for Alzheimer's disease. It's not quite so good for vascular dementia. Interestingly, we have replicated those results largely in our direct studies. So we've been looking at accuracy versus expert adjudication of full free text medical records in a subset of participants in the Lothian region of Scotland looking at relatively small numbers of cases of disease. Summarised on this slide you can see that the positive predictive values that are achieved through a process like this are pretty good, and in particular show the power and accuracy of primary care data which performs every bit as well as hospital data in this regard.

So in terms of future plans, what are we looking at doing? So we're thinking very hard about additional data linkages to enhance the accuracy of what we obtain, to ascertain health outcomes that we don't capture at the moment. So, for example, we are pursuing linkages with microbiology datasets across the UK which will add value and health outcomes we don't currently have and which will enable sub-phenotyping. So for example, some disease registries and audits may give us more detailed sub-phenotype and sub-classification information not currently available. We've been looking at additional web questionnaires, which you've heard about. We are conducting a number of further regional validation studies to assess in detail the accuracy of the linked data for additional conditions and potentially in additional regions of the UK to look at how generalisable our results from the Lothian region of Scotland are. We're thinking about the value of making deep dives into the systems of large hospitals which cover… A relatively few number of hospitals cover a
reasonable number of participants in Biobank for disease sub-phenotyping. So, for example, obtaining information from pathology reports, digitised images and tumour tissue for cancer sub-phenotyping, from free text radiology images for sub-phenotyping of a range of other conditions as well.

Finally, which you'll hear about hopefully in a few minutes, we've been thinking about launching a scalable phenotyping data challenge for two reasons. One is, to try and look at data driven ways of automating this process of phenotyping health outcomes and the other is to engage with an even wider community globally of computer scientists and data driven innovators who can really help us and the research community to address some of these challenges. Thank you very much.

[Applause]