Cathie Sudlow Professor of Neurology and Clinical Epidemiology, University of Edinburgh
June 2017

Thank you very much. So to realise the benefits and study those environmental, genetic and lifestyle determinants of health we need to follow the health of half a million people and we need to find out who gets disease and who doesn't, and who gets what sort of disease. Because all the participants in UK Biobank will develop some sort of ill health problem at some stage; none of us will get away with it. Everybody will be there for informative in some sense but we need to identify which people get which disease so that their data can be used for studies of that particular disease in comparison with people who don't have that particular disease. That's a very big task, it's not something that can be done in a doctor's clinic for example in terms of following up those people, and the very best way to do it is to do it through the NHS data that is being collected all the time. But as you've heard, that's a very tricky thing to do because the data are hard to get hold of for a range of different reasons.

So one of the reasons they're hard to get hold of is that UK Biobank participants are in three different countries. As you can see here there's about 36,000 in Scotland, about nearly 450,000 in England and 21,000 in Wales. Substantial numbers in all three countries albeit that nearly 90 per cent of them are in England. So a follow-up strategy has to take into account the different jurisdictions and the different ways in which NHS data are made available in those three different countries. So in UK Biobank we're now in a position after some years of heavy lifting that we are able to obtain regularly updated information on a wide range of diseases from national NHS data sets in these three countries. We can link to death registers and obtain information on deaths, date and cause of death, and about 14,000 people who were originally recruited to Biobank had already died by 2016. Very sad for them and their families but at least very informative for medical research because we know how they died, what they died of and that's very helpful in terms of studying causes of disease.

We also follow up individuals through national cancer registries and these exist in the three countries as well and we have information from those on both the date and the type of cancer diagnosis, and again, over 79,000 participants had already developed one sort of cancer or another by early 2015. Then we also now have managed to obtain regular linkage updates, not without difficulty, to hospital admissions data, which provides information on both the date and the diagnosis or the procedure that was conducted during a particular hospital in-patient visit. All this information is provided in computerised form so that words, text of diagnoses are converted into codes and we call that coded data, and that's very important for the way in which these data are analysed. So from those three sources, death data, cancer registry data and hospital admissions data, we can identify very, very large numbers of conditions that are developing during the course of follow-up, in participants who, in the main, were healthy at the time of their recruitment to the study. Here are just a few examples for you.

So some common cancers, breast, colorectal and prostate cancer, where we observe in our data set right now that by 2014 several thousand people had developed each of those types of cancer during the course of follow-up. Then using the data that we've observed so far we can predict what we will see this year when we get the data and in some years hence when we get the data that there'll be many thousands more cases of

these common cancers. Of course there are many other cancers that we can retrieve, identify and study in just the same way. I've shown here some examples as well of some common non-cancer conditions although there are again a very large number of other conditions that can be studied in this way. So there are about 25,000 or 26,000 individuals who had diabetes even before they joined the Biobank study but during the course of follow-up, by 2014 6000 new people had developed diabetes and you can see going forward that 13,000 and then 21,000 by 2022 will have developed diabetes, and we can detect them through these hospital admission databases.

Similar numbers, same sort of order of magnitude can be seen for myocardial infarction or heart attack, for stroke and for chronic obstructive pulmonary disease; these are some of the commonest conditions that affect the sort of age group that are being studied in Biobank as they get older. But one of the very important sources of data that there has not been a national solution to is the ability to link to primary care data and all of you know that when you go and see your GP that's the first point of call for any health problem that you have. Many conditions are managed almost exclusively in primary care, or at least for the main part in primary care and that would include things like asthma, diabetes, arthritis and so on; very important conditions and very important for researchers to be able to study and understand those. So again, we have our participants spread around the UK and we need to find a solution. So there are no national providers of primary care data with complete coverage of the country and so we've had to find different solutions for obtaining primary care data in these different countries and after a few wild goose chases we have found the following.

So in Scotland we link to primary care data through a software company called Albasoft, little company based in Inverness, which has its software implanted in all the GP practices round Scotland, and partnering with them allows us to extract relevant data on the relevant participants and we've covered about 27,000 of the 36,000 participants in Scotland in that way. In Wales we partner with colleagues at the University of Swansea and NHS Wales who set up a secure anonymised information linkage service, which allows us to link to primary care records. So all the GPs feed their data into the service and we link to it within that system and that's allowed us to link to almost all the participants in Wales and obtain their primary care data that way. In England it's been tougher to find a solution until we realised that actually there were only three major computer system suppliers of GPs who we could deal with directly. The first of those companies that we developed a relationship with was TPP, The Phoenix Partnership, which has all the data from its practices held in one giant data warehouse so forming the linkage directly with them once we discovered how to do it was relatively straightforward and last year we received data for just under 170,000 of the participants in those practices.

We started to forge similar relationships now with the other two major system suppliers, In Practice Systems, where we work with a software supplier called Apollo as an intermediary, and EMIS and I've no idea what EMIS stands for but someone else might! So EMIS is going to be the last of the companies who we develop a linkage system with and we expect to have data from them for about 250,000 participants probably

sometime in 2018, hopefully earlier rather than later. We do plan to make primary care data available for researchers from a subset of the cohort, probably the TPP cohort because that's the one we've had for the longest and that's the largest in early 2018. So what's the impact of primary care data then on the sorts of disease detection? So just to illustrate that I'm going to show you the first slide that I showed earlier showing some of the numbers of those common cancers and common non-cancer conditions emerging over the course of time.

If we just concentrate on the data, the numbers that we observed in 2014 of those conditions you can see that for cancers they're reliably detected in the registries so it doesn't make very much difference if one adds in primary care data, the numbers don't change very much. But for those not common non-cancer conditions it makes quite a substantial difference particularly for some. So for diabetes, for example, if you look in a hospital admissions data set you only detect about half as many cases as if you include primary care data because many of these people don't get admitted to hospital. If you look for heart attack you don't find very many more in primary care data. It increases from 5000 to 5500 at the same point in time because most people with a heart attack will be admitted to hospital and picked up that way. For stroke we will find about half as many again cases by including primary care data and for chronic obstructive pulmonary disease the numbers more than double, again because most people with chronic chest disease will not necessarily be admitted to hospital but will be looked after in the community.

Now moving from these linked data sources to be able to accurately identify disease caseness, so that people can undertake those comparative studies comparing people who have disease with those who don't, is a non-trivial matter. One of the things we're working very hard on at the moment is trying to combine the information from these various different data sets to create, if you like, off-the-shelf disease status indicators that researchers can use. When researchers open up a data set from the NHS they probably have a quick look and think, oh my goodness, what a mess and close it back down again in a hurry so what we're trying to do is provide a service to the research community to enable them to actually use their scientific imaginations quickly and not have to do all the boring stuff. So we're working with experts in all of these different areas to try and convert these linked data into those disease status indicators and we're doing that in a couple of phases. We're dealing in phase one first with the cancer hospital admissions and death data because we have complete coverage for those and the follow-up is going reasonably well and then in the second phase we're incorporating some additional cancer data, more of which later, and those very valuable primary care data.

So this is my last slide and it's perhaps for the scientists in the audience and for participants who have an interest in or suffer from particular conditions to show you how the rollout of those disease status indicators in the database is going to occur. So in red you can see where the phase one algorithms are going to become available and in blue you can see where we're going to be incorporating into those algorithms the primary care data. So if you look here you can see that already there's very good valuable cancer information direct from the registries, we don't have to do too much there, and if you look on the Biobank data showcase you'll find that there are already these disease status indicators available for myocardial infarction and stroke

and a large number of other diseases is going to be following in the next year or so. Thank you.

[Applause]

[END OF TRANSCRIPT]