DANNY LENNON: Naomi, welcome to the podcast, thank you so much for joining me.

NAOMI ALLEN: Thank you very much for having me.

DANNY LENNON: Yes, I have so much to ask about your work, but I think maybe a good starting point in your role as chief scientist with UK Biobank, can you maybe give some insight into what that actually entails, whether that's a day to day basis, project or project basis, what does your role really look like?

NAOMI ALLEN: So broadly speaking, it's a really diverse role, but the essence of it is really how we can best enable researchers worldwide to make scientific discoveries by enhancing their resource. So my job really is to think about what is the future scientific strategy for the resource, so what is it that researchers need to make those future scientific discoveries happen, whether it's by measuring a whole range of biomarkers, sending out questionnaires to obtain information that we don't currently have, linking all of our half a million participants to medical records to obtain clinical outcomes that we don't currently have – what is it that we can, that researchers really need to enable scientific discoveries to happen – so that's kind of my job in a nutshell.

DANNY LENNON: Maybe for people who are just coming across this or have heard it mentioned recently on the podcast, how is it that you would like to introduce what the UK Biobank is and why it's so important and how it's become so influential for many research teams?

NAOMI ALLEN: Yeah, so UK Biobank is a cohort study of half a million people, men and women across the UK. So we recruited these half a million people aged 40 to 69, about 10 years ago; we invited them along to an assessment center, had a whole range of questions asked about them, loads of physical measures, we took their blood samples and saliva and urine; and they gave consent for us to follow them up over time through linkage to their medical records. And so, we have collected this vast amount of data about their lifestyle, we have their blood samples, so we've collected genetic information about all of these participants and we've followed them up over time, link into the medical records, and we're making this dataset that contains this vast amount of lifestyle genetic clinical data available to researchers worldwide to answer health related questions that are in the public interest. And so, what I think will be unique about UK Biobank is two things. First of all, it's the largest dataset in the world that has such a vast array of health, lifestyle, genetic, imaging clinical data, I mean, that really is unique. And secondly, it's easily available for bona fide scientists to work on. And so that means the global research community has come together and collaborated and has made scientific discoveries on the data that we had no idea that this dataset would even be used to look at that question five years ago. So it's enabled the imagination of scientists worldwide to be fulfilled because you've got this massive dataset to enable scientific discoveries to be made, and I think that's what really is unique about UK Biobank.

DANNY LENNON: That is really a fascinating element that it's kind of leveraging all the sort of creative

ingenuity of various different research groups to come up with these interesting research questions, and now they have this amazing data to work with, and then that all gets fed back in for the wider scientific community. Just out of interest, how many applications do you typically get for access to the data, and what does that kind of application process looks like?

NAOMI ALLEN: Yeah, well, it's growing all the time, and we've really had to try to streamline our processes because we've just been inundated since we opened our doors in 2012. We've now got over 14,000 registered researchers over the world, it just started off with UK academics because we were the ones who knew about UK Biobank, we help to set it up. Now, the majority of researchers are from overseas, a lot from the US, a lot from Europe, increasing researchers from China now, Australia, and across different fields – we've got geneticists, epidemiologists, nutritionists, social scientists, bioinformaticians, mathematicians, all accessing the data, so about 14,000 registered researchers and between 1500 and 2000 research projects that are actively underway. Publications coming out – we can't keep on top of the publications that are come out all of the time on UK Biobank. So it's starting to make a real impact on the research community, especially in the genetic field because it's the largest study to have a really detailed genetic information on a very large population.

DANNY LENNON: Yes, and just maybe to piggy off the back of that, you mentioned that there's this genotyping that's been able to produce this amazing data, and you've alluded to some of the other data that's been able to be collected, can you maybe just give an overview again of what we're looking at, whether that's the imaging, the questionnaires, what type of data have you been able to collect?

NAOMI ALLEN: Yeah, sure. So when people were recruited into the study 10 years or so ago, all half a million

people answered a whole range of questions about their lifestyle, smoking, alcohol, diet, physical activity. We took a whole range of physical measures, blood pressure, weight, height, spirometry, heel density and so on, and then we also took blood samples. So we've done a whole lot of genomic measures, genotyping; we're now doing exome sequencing and whole genome sequencing on half a million. I mean, even five years ago, the thought of the whole genome sequencing on a 100 people would have been considered ambitious and now we're doing it on half a million, and it's kind of routine, it's incredible how the fields moved on. And not only that, in a 100,000 people, we're inviting them back to have imaging. So this is MRI scans done on their brain, their heart, abdomen, a full body DEXA scan of their bones and joints on a 100,000 people which is the largest study in the world to be able to look at internal organs which will help researchers to look at the mechanisms through which genes and lifestyle influence subsequent disease risk. So how do those factors influence the internal body before they develop disease – so really looking at the pathophysiology. And we've also got accelerometer devices on a 100,000 people, so we have objective measures of physical activity, of sleep in one of the largest datasets in the world, and we're also sending out regular online questionnaires to everyone for whom we have email which is about 330,000. And this is primarily to collect more detailed information on exposures, so we've asked much more detailed questions about diet on repeated occasions, so what diet you recall over the last 24 hours on four different occasions, we've asked about occupational history and also about health outcomes that you can't assess through linkage. So things like mental health, pain, sleep, digestive symptoms, conditions which are not easily ascertained through linkage to primary care or hospital records, so self-reported aspects related to quality of life. And of course, along with all of this data, you've also got linkage to medical records. So primary care, hospital data, death, cancer,

which – and now in light of COVID-19, we leverage in these clinical records and making them much more readily available to the research community in order to do COVID-19 work as well. It enables researchers across a whole wide range of disciplines to access the data and be able to see, to tease out the independent effects of different lifestyle factors and genetic effects in relation to disease, because you have all of these measurements on such large numbers.

DANNY LENNON: Yeah, it's pretty amazing. And in addition to that linkage with some of the electronic health records, you also mentioned this huge amount of data you've been able to collect. So one of the issues, I guess, is when there's such a huge amount of data, being able to parse through that and be able to account for all of it can become a challenge although, I suppose, now we're in an area of thanks to technology and AI and the use of algorithms and all these areas in bioinformatics where maybe that's going to become more and more important to tease things out, from your perspective, what do you see that looking like over the next number of years of the integration between not only data collection but how it actually can get used?

NAOMI ALLEN: Yeah, that's a really good point because it's great having all of this information but what the hell do you do with it – it's overload for lots of people. There's 3000 variables in a dataset. It can take two days to download the dataset if you have got a flaky internet connection. And it certainly, it's necessitated people having to have automated pipelines, particularly for the imaging data that was processed really manually before now – when you have a 100,000 imaging scans, you need to develop machine learning algorithms to process those data to create derived phenotypes from the scans, otherwise it will just take years and years and years to develop those phenotypes. And I think having such a large amount of data, it's meant, it's enabled researchers to go from simply testing a hypothesis which was

essentially what we're doing up until about 10 years ago “is x associated with y, let's go and find out,” to, okay, we've got these vast amounts of data, we've got vast amounts of compute power, let the machine do the work. So systems learn from the data, so the more data there is, the better the algorithms become at predicting health outcomes. So now groups are biomathematicians are implementing machine learning algorithms to, for example, they use machine learning on the – we've collected eye images at Baseline, so we've collected a photograph of the retina, and they've used machine learning to identify the likelihood of someone developing diabetic retinopathy, based on the picture of the back of the retina. And if you had done that without machine learning, someone would have actually manually looked at the scan, made an assessment on 100,000 scans, it just wouldn't have been possible. So this use of machine learning and these algorithms can be used to predict health outcomes, and this machine learning algorithm not only could predict things like diabetic retinopathy, it could predict based on the width of your vessels at the back of your retina, it can predict your age within five years, your sex almost 100% of the time, whether you are a smoker or not, your blood pressure, and even your likelihood of developing a major coronary event in the next five or 10 years. So these machine algorithms are pretty good already. So you could see, it's not beyond the realms of possibility that in five years’ times these retinal photograph scans, which are in opticians on the high street now, could be used, the algorithms could be put on the devices to be used to predict your likelihood of developing a coronary event or the vascular events, so it could be used for screening purposes for either eye conditions or heart conditions for the future, so it's really exciting. And with before AI and machine learning, that just wouldn't have even been possible.

DANNY LENNON: I wanted to ask about some of the publications that came from the UK Biobank data – you

mentioned there's been a whole host of them and some very, very important papers and I definitely don't want you to have to start picking favorites or anything like that. But as an example for some people, what are some good examples of publications that you feel have been maybe quite influential?

NAOMI ALLEN: That's a really tricky question, because it largely depends who you ask, because the UK Biobank data is used by many different scientists. I think if you asked an imaging specialist, they come up with an imaging paper; if you ask the nutritionist, they come up with a diet related paper. I think for me, I'm not a geneticist by training, but there's research that's been published and is now being expanded on that's used for genetic data which I think is really starting to change people's mindset about the importance of genetic variation in the risk of developing subsequent diseases. So this first paper is published by the Broad in Harvard a couple of years ago now, and what they did was they looked at, by using the genotyping data that we have, so this is where we've measured genetic variants at 850,000 different places along the genome, and they use that data to predict the risk of common diseases, and they compared that to the risk of developing what we all know and think to be genetic diseases. So for example, some are just diseases that are caused by just a single mutation in a gene that has very large effects, so things like cystic fibrosis, Huntington's disease, familial hypercholesterolemia are all caused by single mutation, one single gene. Whereas most diseases do have a genetic predisposition, there's a family history to heart disease to cancer to dementia but their cause, this family disposition is caused by lots and lots of variation in your genes that each of which has a really small impact on developing your risk, but taken together that cumulative risk can be just as great as what we consider to be genetic diseases. So this group found they developed and validated a genetic risk score, and they

found that 8% of the population in the UK are at least three times – have three times higher risk of developing coronary heart disease if they have, if they are unlucky enough to have all of these little common genetic mutations that increase their risk. And that's just as high a risk as you have with say familial hypercholesterolemia, something that we think of as a genetic disease, which said to me that I was kind of quite wowed by this paper, because this means diseases that we think of as being environmental, heart disease, diabetes, actually there is a large genetic component to these diseases and it's not based on a single mutation having a large effect is that mutations throughout your genome have really small effects, but you group them all together and suddenly there's quite a large component of your risk of developing these lifestyle diseases is actually through your genes. I think what that says to me is that this could have implications for future preventative programs or screening programs.

So you could imagine, going to your GP, having a blood sample taken, they'll measure your genetic profile and they say, okay, well, based on your whole genetic profile, you've got a 20% risk of having a heart disease or diabetes in the next, you know, by the time you're 60 or something. And if that's the case, you've got time to do something about it, you've got time to modify your lifestyle and therefore reduce your absolute risk of developing the disease if you know that you're lucky enough to have these group of genetic variants that make you more predisposed to the disease. So I think, for me, that was quite – it changed my view of the importance of genetics, particularly when looking at how they're causally related to common conditions.

DANNY LENNON: Yeah, that is fascinating, and there's so much within it, and I think one of the other things that pops up from that is that when we start considering the lifestyle interventions that are typically aimed at reducing disease risk, now

when we consider what you've just discussed, that would give us some explanation as to why there is that individual variation in the response to such lifestyle interventions, why some people derive great benefit versus others not, maybe because we just haven't been taking into account this genome wide issue as opposed to focusing...

NAOMI ALLEN: Absolutely. And the classic example is obesity, right, so some people say, oh, I just have to look at a slice of cake, and I put on three kilos. Some people, they respond very differently to weight loss regimes, and that may partly be or largely be because of their genetic makeup. So there's been lots of work done in UK Biobank looking at the genetic predisposition to developing obesity, and it's now clear that some people are just more genetically predisposed to putting on visceral fat or abdominal fat compared to other people. And so that does have implications, if you are one of these individuals that has genetically predisposed to developing obesity, then you can see how you can start to tailor public health advice to different population subgroups if you know their genetic profile. And it also helps, perhaps, it will help to develop treatments as well, so you can start, so, you know big pharma companies are now using UK Biobank in anger because all of this genetic information, they can start to think about drug discovery work, and actually how can we develop drugs that can work for groups of people who have certain genetic variants, say, for example, who can't metabolize a certain drug. So I think this all has implications for both preventative lifestyle messaging and screening purposes, but also at the other end of the spectrum for helping to develop treatments that are more targeted to specific groups of a population with certain genetic profiles, you know, why is it that the people respond to cheomotheray and others don't. It's a bit of a black box, nobody really knows why. And if we knew more about their genetic profiles, we could perhaps target those treatments more effectively.

DANNY LENNON: Yeah, and it does seem that the promise of looking at a genetic profile is where it's at because we've had so many cases in the past where a reductionist view of trying to focus on one gene, and does this gene cause obesity obviously gets dismissed rightly so in scientific circles despite maybe mainstream media jumping on some sort of this is the gene that causes obesity whereas most of the work doesn't actually show that, and maybe starts to shift to this looking at the whole genome; and then beyond that may give more of a footing to shift some of the stigma away that's usually attached to something like obesity in particular, where it's seen as a solely this lifestyle thing.

NAOMI ALLEN: Yeah, you're absolutely right. And I think trying to move away from only the FTO gene is important in obesity. Well, actually it's lots of different variants across your whole genome. We don't know what most of them do. We don't know whether they're functional or not, that's work to do for the future. But we can generate these genetic risk scores that are much better predictors of whether you'll develop obesity or other health conditions over your lifetime based on the variation across your whole genome, and that's much more powerful than what we have been doing up until now which is focusing on single genes which we think may be involved because we know their function.

DANNY LENNON: There are so many different areas we could focus on, and I think I would like to explore some of the data around cancer, especially, given some of your work in the past. From a broad view overall, of course, we could have to look at each cancer individually, but what type of data have you collected on cancer specifically and where have you been able to pull that data from?

NAOMI ALLEN: Yeah, so we collect routine data from cancer registries which tell us when someone was diagnosed with cancer and the type of cancer

that they have. And we're also just now processing the data, it's not yet publicly available, but we hope to make it available in the next six months is data on the stage and grade of cancer diagnosis so you can tell how advanced it is and also cancer treatment. So there's lots of data that exists about whether people have had chemotherapy, radiotherapy, the drug name, the dose, the regimen, all of that data will be available for researchers in the next six months or so. And that will really help researchers to be able to look at cancer survival, because without having information on how aggressive was the disease when it was diagnosed and what the treatment regimen was for that individual it's very difficult to look at what things impact cancer survival. And there's lots of research going on in the UK that is really focused on this question at the moment, because it's well known that the UK, in particular, the cancer survival statistics for some cancers are not as high as they are in other European countries. So there's lots of research and lots of kind of pressure to try and understand why is this the case, why is this the case, and what are the factors that – is there geographical variation in cancer survival, is it different across different ethnic groups, why is that, is there some genetic factors determining cancer survival, is it lifestyle factors but you really need information on disease aggressiveness and treatment in order to be able to really get to the bottom of that. So that's why these particular data that we're processing now as a priority will be hugely impressed to be able to answer those types of questions.

DANNY LENNON: Yeah, and I guess with cancer being so unbelievably complex if we're just to look at a diagnosis of a certain type of cancer, like you say, that may not give us all that much information, given that even within one type of cancer we could have tumors that act very differently, and so knowing some of those things around, how they're treated would give much more insight, I guess, but it just speaks to the complexity of even thinking of cancer as

one thing as probably a misnomer in many ways.

NAOMI ALLEN: Yeah, it is, and even – one of the beauty of you UK Biobank, because it's so large with half a million people, even when you're looking at a cancer like ovarian cancer, off the top of my head, I can't remember how many cases of ovarian cancer but, let's say, it's 2000, something like that, it'll be of that order of magnitude, we've got to say 2000 women who have been diagnosed with ovarian cancer since they joined the study, ovarian cancer, there's different types of ovarian cancer, and depending on the type of cell it originated from and so on. So the question really becomes: do these different types of cancer have the same etiology? So are they caused by the same factors? It's the same for breast cancer. We know that some are estrogen receptor positive and some are negative, and it's thought that they have got different causes, so different lifestyle factors affect estrogen positive cancers compared with those who are estrogen negative. So you really need the information on the pathology and the histology of the cancers in order to be able to really pin down what are the cause of lifestyle and genetic factors for these different sub classifications of cancer.

DANNY LENNON: On the podcast before we've talked about some groups that are hypothesizing about certain dietary interventions that may be used as an adjunct therapy and the ketogenic diet gets often, you suggested as one that could be used. And then that unfortunately gets broadly thought of as, oh the ketogenic diet is good for cancer. And then when you look at different types of cancers, you see that tumors respond very differently and, in fact, some tumors could grow much more aggressively in the context of someone being on a ketogenic diet. And so, it just kind of goes to highlight exactly what you say understanding what is driving the cause of a specific type of cancer, and knowing that before taking any implications from what certain interventions may mean is critical.

NAOMI ALLEN: Yeah, I agree, and I think having – the future really, particularly for this particular dataset, is to be able to look at gene environment interactions for cancer. So given you've got a certain genetic profile, does certain lifestyle factors make you more or less likely to develop a certain type of cancer compared with if you haven't got that genetic profile, and is there some kind of interaction there or are they actually independent. So you really need very large numbers to be able to do that kind of work robustly anyway; and we've already got 4000 cases of prostate cancer, the same number of cases of breast cancer, so we can now start to do that type of research and we're not reliant on bringing lots of different studies together in a pooled meta-analysis to be able to answer that question, each of which have measured things differently, ascertained their cases differently, because the study is so large all the measures are standardized, you can do that robust research within one study population. What I will say though is the UK Biobank are a healthy bunch compared to the general population. So they tend to be a bit wealthier, they tend to be a bit healthier. So all in all, the overall death rates are lower than the general population. And lung cancer rates, for example, about half of the general population because most of the cohort don't smoke. So you kind of need to bear that in mind when interpreting results, but just because the UK Biobank cohort is not representative of the population, it doesn't necessarily mean that any associations that you find between an exposure and an outcome are not generalizable to the population at large.

DANNY LENNON: Yeah, that's interesting, and that was something I was going to ask about how those rates compare between the Biobank data versus the overall UK population. So that’s interesting that both incidence rates and then also the likelihood of succumbing to a certain disease probably change for both those metrics. With some of the trials that have been or some of the

publications, I should say, that have been done specifically around cancer using the data, are there any of those that jump to mind as ones that you think were particularly informative in the field or even that you found particularly interesting in some way?

NAOMI ALLEN: So I would say, it's still a little bit premature in the cancer fields, I mean, it we've got very large numbers for prostate, breast, colorectal, the few most common cancers. But for the rarer types of cancers, I think, we still got to wait a couple of years until there's enough in incident cases of cancer to really do rigorous, rigorous research. We're just actually about to update the cancer statistics up until 2019, so there will be lots more cases available in the dataset. But I think, for me, so far, there's been quite a lot of research – and I know you had a previous podcast about this – quite a lot of research looking at diet in relation to cancer. And we've been able to assess diet both kind of through a short food frequency questionnaire at Baseline but also through a much more detailed online questionnaire, and that's enabled us to estimate macronutrient intake and food groups, and particularly findings related to processed meat intake in relation to bowel cancer. We are kind of, you know, there were studies that have shown that before, but I think the UK Biobank was one of the largest studies and really cemented that finding. I think everybody, there's a large consensus now, but that observational finding is robust based on that work. And also, there's been some work done looking at full blood count and hematological parameters relating between people who are vegetarians and meat eaters and vegans and so on and finding that, for example, vegans were more likely to have anemia than meat eaters. Again, you might think, well, that's perhaps to be expected, but this is the only study that has been able to show that empirically in such large numbers. So while it's still kind of pretty basic cross-sectional research, it nonetheless is a really robust observational finding, and that means

you can really build on that work to then think about, okay, so how are we going to address this problem, what can we do, what are the implications of being vegan or vegetarian on future health outcomes. So I think the best is yet to come for cancer.

DANNY LENNON: Yeah, and I think more broadly when you discuss diet and its use from the Biobank data being able to use questionnaires but also then follow up repeat measures of diet which is something that could be incredibly important within epidemiology.

NAOMI ALLEN: Yeah, I think one of the important points there is the value of repeat measures for diet, nutritional epidemiology has been rightly criticized in the past for rather crude measures of diet. But if you have repeated 24-hour recalls over time, you can start to take both measurement error into account but also changes within an individual over time into account to get much more precise estimates of the risk associated with various dietary factors and health outcomes. So the value of repeating dietary assessment over time in a large population goes quite some way to addressing that criticism.

DANNY LENNON: With looking at diet-disease relationships beyond cancer, what have been some of the other areas that have been looked at by any of the different researchers that have accessed the data?

NAOMI ALLEN: So there's been quite a lot of research looking at diet in relation to heart disease and fat intake and meat intake and so on; and also in relation to various eye disorders as well, so fruit and vegetable intake reducing the risk of cataracts and so on. Again, all of these are kind of building on hypotheses that have been suggested from previous studies done over the years, and the UK Biobank data is kind of cementing those findings. So I think the research being done particularly around in the diet space in UK Biobank, I think it would

really benefit from now being incorporating some of the genetic data, incorporating some of the imaging data, so we can start to understand the mechanisms through which diet/obesity impact disease risk by starting to look at impact on abdominal measures, effects on the liver, the kidney, and so on. And the beauty of these measures is that you can objectively measure the distribution of body fat, so you can quantify visceral fat versus abdominal subcutaneous abdominal fat. Up until now we've relied on crude measures of BMI waist-hip ratio as proxy measures for general and central obesity. With the imaging data, you've got direct measures of visceral fat and genuine fat. And I think having that data on a 100,000 people will really start to pinpoint what aspects of body fatness really determine cancer risk, for example. For me, that's a really exciting area for future research that I'm very interested in.

DANNY LENNON: And speaking of kind of future research beyond that example you just gave, are there any other specific research questions that you are excited at the prospect of people being able to work out or test using the data?

NAOMI ALLEN: Oh there are so many. Well, I think, we've talked quite a lot about genetic risk scores in relation to common diseases, and I think being able to incorporate genetic risk scores into other risk scores that contain lifestyle factors such as smoking, BMI, alcohol consumption, age, gender plus genetics, the combination of that should really start to be able to predict risks of common diseases with greater precision than we've got at the moment. So I think that's an area of research that the UK Biobank can enable that type of research to happen. Also, the gene environment interactions which we've touched on already, whether there's kind of synergistic effects between genetics and lifestyle factors on the impact of disease, that's an area which I think will – I'm excited to see what happens there. But also with the imaging data, and we're also looking into whether the – the technology is

there now where we can measure thousands of circulating metabolites, thousands of circulating proteins, so metabolomics, proteomics at scale on everybody. So what are the mechanisms through which diet and obesity influence risk, how does it change your circulation levels of metabolites and proteins in the body, you can start to unravel what's going on between lifestyle factors and disease risks through actually mechanistic pathways. And I think that is really, it's a real black box at the moment, and I think UK Biobank could really start to shine a light on that.

DANNY LENNON: Wonderful, so much exciting stuff going on. Now, before I get to my very final question, for people who are interested in finding out more about the UK Biobank or more about you and your work specifically, is there anywhere on the internet that you can direct their attention?

NAOMI ALLEN: Yeah, sure. So we have a Biobank website which is, well, you could just Google UK Biobank, you'll find it, or it's biobank.ac.uk – lots of information about the study. And also we have a data showcase, so if you're interested in actually accessing the data and doing research yourself, you can see exactly all the variables we have, the type of data we have, you can see the distributions of each and every data field. So there's lots of information on our website. And yeah, I'm very, very happy for people to contact me directly and you can find me on the Big Data Institute's website at Oxford University.

DANNY LENNON: Brilliant. And so with that, if you could advise people to do one thing each day that would have a positive impact on any area of their life, what might that one thing be?

NAOMI ALLEN: Let me say this, because it's based on, it's what I should do more, and I tell myself to do this every day and I fail, is to sleep more. I only say this because I've been talking quite a lot to some sleep experts. We are developing an online survey to find out more about sleep

conditions, and we really need to know more about the effects of lack of sleep on health conditions, particularly on obesity but also chronic health conditions. We know already that night workers have increased risk of certain conditions. Why's that? Is that melatonin or is that an indirect effect of sleep deprivation? So it's an area that I'm getting more and more interested in, and so I think my recommendation would be try and get a bit more sleep because that affects your well-being and has indirect effects on health outcomes I'm sure, but there's much more research to be done on that.

DANNY LENNON: And so with that Naomi let me say thank you so much for, first of all, taking the time out to do this and for the great conversation but also more broadly for the work that you do and that you continue to do, it's very much appreciated, so thank you for being on the podcast.

NAOMI ALLEN: Been a pleasure.