Genetic and protein data in understanding society

Meena Kumari: OK, so my name's Meena Kumari and I'm a co-investigator of the Understanding Society dataset and I look after the health and biomarker data at Understanding Society. So today we're going to focus on the genetic and protein data that are available in the dataset. But I'll give you a background, I'll start off and just give you, I don't know if everybody knows everybody here knows the Understanding Society dataset. So the first half of my, I'll be talking about half an hour, and the first half my talk I'll be talking about the data that are available in Understanding Society and then I'll talk a little bit about the biological data that are available, so this is an overview, so the health and biological content, and then I'll just do a brief discussion about terms, because there are lots of there's lots of jargon and a lot of terms used with the biological data. And it's sort of changing over time and so I'll just give you a definition, so that when you, if you come to these data types or these datasets or this literature, you have an idea of what people are talking about, and then just examples of how the data is being used in ways that are interesting for social scientists.

So Understanding Society is the UK Household Longitudinal Study. It's a survey of individuals and it's representative of the UK population. And we interview everybody in the household so it's a household panel study. We select, we randomly select, households across the UK and we collect information from everybody in the household, so that's adults and children and there's a core sample of participants and we interview everybody annually. It's a basic design, it's similar to a number of the international household panel surveys, PSID, HILDA. You can do lots of comparative analyses, I think the strength of Understanding Society is that we do have the biomarker data that I'm going to talk to you about, that a lot of other studies don't, so that kind of makes us unique and special. The study started in 2009 and it builds on the British Household Panel Survey and there's lots of different types of information that are collected. So what people are experiencing, attitudes and identities, wellbeing. And obviously with the interviewing participants annually we're getting an idea of how these things respond to policy changes, key events in people's lives, a big event has obviously just happened to all of us in terms of pandemic, and so we can see what was happening to people before, before the pandemic, we'll be able to interview people afterwards, we'll be able to see what happened to them afterwards, so there's lots and lots of research possibilities, measurement of change over time and because we're interviewing everybody you can examine what's happening to people across the life course so it's a really important survey to think about how policy environments are changing what's happening to people, how what's happening to people in the face of changing kind of norms and environments so that's sort of kind of one of the strengths of the survey. It's composed of four kind of main parts, there's four main parts of the survey. So the general population sample which is a sampling of the of the general population, so it's 30,000 households and we've got families across the across the UK, as I said. There's an ethnic minority boost there's actually two minority boosts in the survey so you've got an ethnic minority boost which is sort of established ethnic groups, really: India, Pakistani, Bangladeshi, Caribbean, African ethnic minorities and then recently we've had a migrant boost which was the Eastern European migrants, so the newer groups that came and the more established groups are boosted in the survey and the survey incorporated the earlier British Household Panel Study and we have biological data from the general population sample and also British Household Panel Survey. There's a testbed of participants, that we do methodological development and researching called the Innovation Panel, and we all have some biological information in that panel now because in IP12 I did an experiment looking at how to collect biomarker data from different, using a different mode, so nurse interview and web-based, so there's biological data in three of these sub-groups of Understanding Society.

This is a long schema of the of the of the dataset. So as you can see, the red square, red rectangles are where we collected biomarker data, so we collected in wave two from the UK household, from the general population sample of Understanding Society, so we had one wave from that set of participants before we collected biomarker data, and we collected biomarker data from the BHPS sample in wave three of Understanding Society so we had 20 waves of of data before we did that biomarker collection so there's lots of possibilities here in terms of doing analysis with the, looking at biomarkers, with the dataset. As I said, we measure the richness of people's lives as core data that we collect from participants every year, so education, employment, health and wellbeing, civics, income lots of core things that we're collecting all the time and then a rotating set of topics and we do that because it's impossible to collect everything at every wave and we have to think about how much burden we're putting on the participants in terms of their data collection, so we interview them for about an hour and collect all these different bits of information. The health and wellbeing, the health elements of the study are collected in two different ways really, one from questionnaire, so asking people about their health at wave two and three we have information on things like prescriptions, at every way we ask about, we ask a whole bunch of chronic conditions, we ask at every wave we're asking or administering a questionnaire on health, called the general health questionnaire, at every wave we ask about mental health and physical health functioning. And then, as I said at waves two and three, we administered a nurse visit so where we collected, a number of different objective data, so heights and weights, where a nurse went to the participant's home and did a set of biological measures from the participants.

So this is just more detail about the sorts of things that we've collected from the participants in terms of their health: the conditions, the SF-12, the GHQ-12, health behaviours, and we have tried to consent, we have consented the participants to data linkage, although we've not been tremendously successful in actually doing that linkage. So we're trying to do that at the moment, we have data linked to the Scottish survey, sorry to the Scottish health data, and we should be having data link to the Welsh data as well we just don't have English linkage at the moment and we're trying to do that. And this is the non-health content. I won't read it out, there's a lot here and we're recording this so you can come and have a look at all the different things that we've measured in the dataset. And, as I mentioned, last year, the year before last year, the pandemic came upon us and we were quite nimble in Understanding Society in that we managed to administer a set of questions, I think we were the first longitudinal study, national longitudinal study, to administer and deposit a dataset on, about the pandemic, so lots of information about symptoms, and then all of the things that we thought might be, have resonance or related to what was happening to us last April, in terms of lockdown, so there's lots of things here that we've measured and we've measured again and again, so we were initially measuring monthly so April, May, June and July and then every other month: September, November, January this year and then March this year, and we just, we administered another survey in July and we've just administered another one, which will be our last one as a lot of the content has migrated into the main data collection, so that we don't need to do these extra covid collections.

So, in terms of the long term content, the questions that, we repeat lots of questions, we rotate modules, we are experimenting with using something called "event-triggered" data collection, so if something happens to a participant, an example of this is the pandemic, so the pandemic happened, and we were nimble and we could ask lots of sort of sort-of focussed questions about that. We're thinking about doing that for other things, you know if you get married, if you have a baby, if you retire, so thinking through whether we can do these sort of event-triggered collections and age-triggered collections that are focused on the particular event that's happening to you. So this is the long term content, you can see how detailed it is and also, but also how it can mean that not everything is asked at every wave and that you, you kind of need to know the data set and go and have a look and see what's asked when and how you could think about how to analyze the information that we have in the dataset.

This is what went into the nurse assessment. We collected heights and weights, respiratory function, blood pressure and then we collected our blood sample and it's the data from the blood sample really that we're going to focus on in the next hour or so and, in that we measure lots of, a number of analytes and we collected DNA and we were sort of interested in how we use that information and whether that information can be used in a social science context. So we were quite careful about what we measured, we sort of thought about what would be most useful, so our criteria were around environmental factors and whether they might be associated with the biomarkers that we wanted to measure, whether they were going to help us understand the pathways by which the social and health might be linked, whether the health outcomes that we were trying to capture affects a reasonable proportion of the general population, and you know there's a reasonable prevalence in terms of the health outcomes and also we're kind of limited by the, because a nurse goes into the participant's home and the blood sample is sent through the post, which means that not everything can be measured, because some things aren't robust to that sort of collection. And so we were looking at core markers for disease and things that we thought that would be useful, either on their own, or in combination, to help us understand sort of how the, how the environment gets under the skin. So those are the sorts of pathways we're thinking about. And this is a list of those the analytes that are already available in the dataset so you can go to the archive now, and you can have access to these data, they're already there, they were measured from the sample, I think they were deposited a few years ago now, so and there's lots and lots of examples of people using these data and doing lots of interesting analyses with them.

So, you know, why would a social scientist be interested in biological data? As I mentioned, we tend, we focus on trying to understand the pathways by which social social circumstances and health are linked. So really understanding, you know, what is it about the environment that might be impacting our health, which things, sort of pollution might be working down one pathway, occupation might be doing something else, so thinking through all of those different processes. We're also thinking about sub-clinical markers of disease and objective markers of disease, so I can give you an example of how we did that now, but also thinking about those sort of the "clinical iceberg", people will often walk around, sort of on the pathway to disease and and not know it. A few years ago, you would, half of the people who had diabetes didn't know it, that's not true now because it became a focus of clinical care, as it were, and so there's less, there's fewer people don't know that they've got diabetes, but if you don't know it you're walking around with it for a little while and it's something that's coming down the road so that's quite important in terms of understanding population level health and identifying populations at risk. And the other thing that we're particularly interested in or thinking about was that a lot of social science is different to other sort of disciplines in that it's open science and again that's changed over time. But it wasn't and it still actually isn't quite normative to be depositing data or making it available and having lots of biomarkers in in a social survey means that it's an example of a more sort of open science in the sense that anyone can go and get these data, as I mentioned, the data that I've talked about so far is available in the archive and you can go and do your own analysis, it's not dependent on me. And you know I'm sure people have lots of interesting ideas and things that I would never think of and so that's a real strength of the dataset and this approach.

So today we're talking about "omics" so we talked about genomics, proteomics, epigenomics, so there's lots lots of "omics" that have become a sort of a term that people have started to use and it's a term that's used for a slightly different approach to the way that we do these things, and has been done in the past, so it's a totality approach. So you hear about people doing GWAS and GWASs are genome-wide association studies and in a genome-wide association study what you've done is you've looked across the whole genome and looked at it and you're interested in looking across the whole genome and your phenotype or your outcome of interest. And so there's a whole branch of science now that's moved kind of down that direction, they have genome-wide scans, they have epigenome-wide scans and these data are available in Understanding Society. And what we have done, and there are going to be talks after me that go into this into much more detail than I'm going to do just now, but this is a sort of introduction, is that we have measured signals across the entire genome, so these things, called single nucleotide polymorphisms, we'll explain that later, but we have data that has measured 14 million of them across the genome and you can look to see how that's associated with your phenotype of interest or your variable of interest. We have measured methylation, which is a measure of something called epigenomics and again we've measured 850,000 sites across the genome, so you can look at that, with your outcome of interest, and we also are going to make available proteomics data so we've measured 184 proteins - that's actually a very small sub-sample of the potential number of proteins that are available in biology, which is thousands and thousands, so a hundred thousand protein probably measurable or available in biology - but we've measured 184 and you can look at those, we'll make those available soon, and you can look at those and your phenotype of interest. And so people put these things together and talk about multi-omics, integrated omics in terms of these types of data.

There are other terms that you might come across in this, in this new omic world, and so people do something called "exposome" approaches and that's really everything that's not genetics or genomics in terms of your exposure, so everything you've ever been exposed to is the exposome and when you go to that literature at the moment it appears like it's everything that you've ever been exposed to in terms of chemicals really so the pollution literature often talks about it being the exposome and it's really all the different chemicals that you might have been exposed to. I think in the long run, if we really are serious about everything that you've ever been exposed to that's not genomics a dataset like Understanding Society becomes really important because we've measured lots of things really well and we've got everything that you know we've got an idea of kind of life course exposure to things. And then, I came across recently this new term "culturomics" which I don't actually know very much about but it's the use of big data to understand human behaviour and culture and culture and again it's sort of trying to capture the entirety of an exposure, having said - it's a new term - it's actually used in a different setting as well: if you're a microbiologist or a lab scientist, culturomics is often the things the sort of microbes that you have ever been exposed to so sometimes new terms come in to use and they they are or are not associated, you know, take on new meaning, so it's all really good fun actually, sort of these new worlds.

So I'm going to quickly just do a one-pager on genetics - we're going to have a much more detailed and better explanation of genetics in the next presentation, but just as an introduction, so that you've got, there are lots of terms, as I said, so we've measured SNPs in Understanding Society and methylation in Understanding Society, and so what that is, in my one-pager of genetics, so you know that DNA - I hope you do - is found in your cells in the nucleus of your cells and it's kind of bundled together in chromosomes in the nucleus so you've got about six foot of DNA in every cell in your body and it's all wound up very tightly in these chromosomes and it's made up of base pairs, so these chemicals called cytosine, guanine, adenine and thymine. And together that's sort of your "alphabet", as it were, and you read that and they're sort of packaged up in genes and those genes make proteins and that's really what we're going to be talking about today. And methylation, which is the sort of a way of dialing up or dialing down how much of the genes are expressed to make the proteins are chemicals that come in and sit between these Cs and Gs so the cytosine and guanine, to impact how much they can, the machinery of the cell can come along and influence the gene expression and the protein that you make. So I would say across this diagram we have measured the Cs and Ts, the CGs and As which are the SNPs, so those might be the same or different between people and impact what protein is being made. We've measured methylation which impacts how much of the protein is being made and we've measured 184 proteins so that's sort of what we're going to be talking about today.

As I said the way that we use biomarker data is to sort of think about the pathways by which the environment and health are associated. And in the past we've gone down this sort of "candidate" approach and that sort of contrasts with this "omic" approach, so we have written lots of papers using this sort of candidate approach, so if you're interested in, if you think inflammatory markers for example are important for mediating the association between the social environment and health, you know, we have already had two proteins that index inflammation and we've written about how they're socially patterned in the past. And so, this is a paper that was written by Apostolos Davillas, Michaela Benzeval and myself in 2017 describing how, in the different age groups, you can see a difference, by educational attainment or income, in people who have higher education or no qualifications and that the difference in inflammatory markers between those groups emerges in our participants who were in their 30s and remains there until they are in their 80s, and so this diagram sort of gives us a description of two inflammatory markers. Other people at the same time were trying to you know, maybe we haven't measured the right inflammatory marker and have looked at a wider set of inflammatory markers this paper by Castagné measured 28 inflammatory markers and looked at occupational differences in those inflammatory markers and created a score and to help us understand which inflammatory markers might be most pertinent to social differences in health and so there's a sort of - you can see a sort of move from picking out one or two things that might be interesting, to looking at things in more detail, to looking at trying to look at everything and seeing what comes out. And here again, these two papers decided that inflammation might be quite interesting and looked at methylation of the genes regulating inflammation and again they picked out a whole bunch of sites across the genome that potentially are associated with inflammation and look to see whether, how the genes associated with inflammation are methylated and whether that, whether that methylation was associated with social position. These two papers reported that they were associated with social position one arguing that it was early life social position that mattered for methylation of these genes and the other one that it was later life that was important, sort of related to methylation of these genes, so even here you've got something quite complicated, in that, you know, different parts of the [inaudible] where it's most important to look. So there's a sort of an idea that a different way of doing this is to actually just look across the entire genome and look to see if we can do it the other way around, so here we've got an example where people have looked at the entire genome and looked to see whether its associated with, methylation might be associated with age which is the top panel there or with smoking. So here what the group did, so Steve Horvath is the sort of lead in looking at age and methylation and what he did was he, what are the conventions in methylation and epigenetic work is that you do deposit your data so but with very little additional information so there are a lot of places you can go to get methylation data [inaudible] age and sex with the methylation data and what he did was just use that information to see how methylation is changing with age and he created a biomarker of age, which has been used actually recently quite extensively as an objective marker of age and we have those data in Understanding Society. In the in the bottom panel there people have looked to see if we can use methylation, how methylation is associated with smoking so again, we just looked down the entire genome and they used 450,000 sites across the entire genome and looked to see if it was associated with smoking and then people have subsequently used the scored that have come from that to see if we can use biomarkers of smoking.

We have this, we have the age biomarkers in Understanding Society and we're making those available, we're going to deposit those in the archive soon. We looked to see if Steve Horvath's - there's a few of these biomarkers of aging - we looked to see if Steve Horvath's biomarker of aging and another one called Hannum biomarker of aging where it actually associated with aging our dataset, and you can see in the graphs just at the top there that they are associated with age in our dataset but not in a parallel way, so our biomarkers of smoking were overestimating age in our youngest age groups underestimating age in our oldest age groups, suggesting that our younger participants were biologically older than their chronological age would predict and our oldest participants were biologically younger and average than than their chronological age would predict. And so that's quite interesting because these biomarkers of age were being used by social epidemiologists to look to see how, you know, how they varied by whether you are biologically different to your chronological age is associated with for example social position and it, for, if you're using education to do that, then education means different things at different age groups, and if this difference in biological age and chronological age is age-patterned that's quite important, and so we published a paper suggesting that you needed to be careful about how you're treating age in these analyses. And what we saw was that, for the social class was was associated with being, so disadvantage in early life was associated with being biologically older than your chronological age so you're aging more quickly biologically if you'd experienced disadvantage in early life. We're also doing some work on how the biomarkers of smoking are functioning in Understanding Society, so this is, at the bottom here some unpublished work from a student, a Soc-B student in our program: he's looking at how biomarkers of smoking are associated with smoking, self-reported smoking in our dataset. And whether that, what are the predictors of mismatch is, and you can see here that our data are suggesting that if you have lots of education, although fewer people report smoking, there is much more mismatch between the biomarker of smoking and self report, so it looks like more people are telling us that they, fewer people are telling us that they smoke than actually are, if you're more educated, so we're sort of looking into that, so thinking about what that might mean.

And so what we're going to talk about next is this sort of omic approach to analysis and what has been done in the social science setting with that, so there have been these sort of genome-wide association studies of things like educational attainment, risk behaviours, a number of, so smoking, alcohol intake, lots of behaviours and people have used those and created these sort of scores, and we use these scores - these scores have been used in approaches to analysis so something called Mendelian Randomization which uses these scores as "instruments" in IV-type analyses to see if we can kind of "instrument" the thing that we're interested in. And then we've, there's also been newer work looking at epigenome-wide association of these things, so educational attainment, trying to understand sort of what biology is being uncovered by these epigenome-wide approaches to social science variables. So the idea about, with this omic approach is that what you're going to do is you're going to look down the whole genome and and kind of uncover new biology or support the candidate approach if it's there, and really that's what you're trying to do. What generally tends to happen is that you're kind of going down the genome and you find that some things, as I say, for example, in this particular paper looking at educational attainment, they found that lots of sites were associated with educational attainment - nothing is associated individually very much and even when you add up everything that's associated with educational attainment it isn't predictive of educational attainment but the scores that you get from doing this work kind of can be used, they're predictive enough to be able to be used as an instrument and people have tried to use those properties of the scores that come out from from this sort of work - I've been going on for too long, I will stop soon, [inaudible] - so what we're going to do next is talk about the biology, so go over again the biology, that I kind of went through very quickly in my presentation. And then we'll talk about the sort of methodological approaches to the data and think about some examples and, finally, we'll present some, what the datasets are and how you access them.