The second part of this talk is about some of the data quality issues researchers should keep in mind when they are analysing biosocial data. I'll be talking about two kinds of issues in particular. One is about the mode or collection condition whether that's important and whether that should be taken to account when analysing biomarkers. And I'll also be talking about some of the quality control or QC processes that should be looked at at least by social science researchers when they're looking at biological data sets. A number of bio social studies tend to use two kinds of methods to collect biomarker data. The gold standard is the clinical lab collection that's because it's very standardized it's a very standardized process participants are invited to come to the clinic and so the blood samples are collected and stored and processed immediately and so because that's a very controlled environment the biomarkers that are obtained are thought to be of a very high quality standard and that's the way how a number of studies in the UK have collected their biomarker. The Avon longitudinal study for example ALSPAC. The 1946 birth cohort study when the participants were age 63 Hertfordshire cohort study and the Whitehall two civil servants study.

The other way of collecting biomarkers are blood based biomarkers are when participants are visited at home and their blood samples are collected and then posted to a laboratory and that's the case with understanding society, Southampton women's study, the 1946 birth cohort study when the parcels were aged 53 the health survey for England and the English Longitudinal Study of Aging. Now when people are invited into a clinic condition as I said it's a very standardized environment so that the temperature is standardized for example the collection conditions are standardized and a blood sample is usually drawn through venepuncture so they draw blood from through a persons veins and that blood sample is immediately processed and then and either stored in the freezer or the blood analytes are measured then and there. In contrast in population surveys what happens is that there's a delay in the processing of that blood sample and the storage of that sample.

And here we've got some examples of the different conditions in which blood samples are taken and stored. In the top left-hand corner we've got a nurse in a clinic setting and as I said that is a very controlled environment and so the blood that the nurse draws is immediately stored and processed in very high quality ways whereas on the top right hand corner we've got a nurse taking blood pressure but could be taking blood as well from a participant at home and there the environmental factors are much more variable for example the room temperature could be could be highly

variable from one home visit to the next. The time of day is particularly important as well so in a home visit the nurse may not have complete control of what the participant did or just prior to the their visit. So the person could for example have smoked or I had a lot of food which could influence the levels of biomarkers. Whereas in a clinic visit there's to some extent some control over the person's activity the respondents activities just before the samples are taken. As these the blood samples that are taken by the nursing home visits are usually stored in particular ways and they can be sent by post for example in a jiffy bag and the process of posting the blood samples the environmental conditions in which these blood samples are then exposed to are as you might imagine quite different from the setting in a clinic situation where the blood samples are stored immediately in freezers. And in the bottom right hand conditioned bottom right hand part of the slide you see some pictures of other conditions that affect both home visits as well as clinic blood sample collections which is that the whether it's a weekday or weekend or what month of the year all of these factors do tend to affect some of the blood based biomarkers so it's important to keep these considerations in mind when analyzing biosocial data.

In the top graph I've shown you the distribution of the times in which nurses visit the people at home in Understanding Society and we see that it's a bimodal distribution we see that nurses tend to visit people just after 10 o'clock or if people are working they tend to visit people at home around 6:00 or 7:00 p.m. at night. So you can imagine that if these are the times in which blood samples are taken and time of day has an effect on particular analytes then we really need to be considering what time people took their biological samples. In the bottom picture we've got a diurnal distribution of the stress hormone cortisol and in general people have a have a very marked diurnal pattern so as people get out there their cortisol levels shoot up and then all the rest of the day that cortisol levels come down so you can imagine a nurse visiting somebody at home collecting cortisol data in the morning it's likely to have very different levels of the stress hormone cortisol compared to nurses doing something else later in the evening.

I'd also like to talk about some of the quality control issues that researchers should be aware of when looking at the biomarker data and that is largely down to the labs that process these blood based biomarkers and they're divided into internal and external quality control processes. Some of the biomarkers have impossible values so for example some biomarkers like height and weight are there they're likely to you know you can imagine

some if you have somebody in your in your data set that is 10 meters tall or a thousand kilograms in weight you can imagine you know that it's going to be hard to anonymize that person because you know they're going to be a 1 in a completely unique individual. So you can pretty much rule out that there are impossible values so it's good to for you know look at the distribution of off of your biomarkers to see whether there are some impossible values and you know treat them as outliers. But independently of that when the biomarkers the blood based biomarkers are processed in a laboratory what the laboratory does is that it tests when it goes through the procedures that derives the blood based analytes it repeats this on another day and hopefully there's a very strong correlation between that analytes that they get on one day compared to another day so that's called a intra-assay coefficient of variation ideally we want there to be a very small amount of variation so less than 5% is within acceptable limits so that's comparing how one particular biomarker within a lab compares to the same biomarker when it's when it's processed on another day. But the external quality control measures are comparing how the lab does in relation to other labs in processing the same analyte and that's measured through the standard deviation index which is a measure of total error in analyzing a particular biomarker in comparison with all the range of labs that have analysed that particular analyte. And so once again we're trying to get at low values of the standard deviation index we want to have the analyte that is measured by the this particular lab to be close to the overall levels that are measured by all the all the labs or score between below one standard deviation index is generally very good.

And also I'd like to talk about specific biomarkers so I've so far been talking about biomarkers in general but you when you get at sort of specific blood based biomarker analytes we should be keeping in mind that each one of them has a different meaning has a different significance so this slide and the next slide looks at one particular measure called c-reactive protein CRP for short and it's a measure of systemic inflammation. And usually somebody that has values of CRP between 3 and 10 milligrams per litter is denoted to have systemic inflammation. However somebody could have greater than 10 mg per litter CRP values which denotes a current or recent infection so the meaning of high levels of CRP is completely different when it's when it's over 10 as compared to when it's between 3 & 10. So when it's between 3 and 10 it so it's considered a measure of cardiovascular risk when it's over 10 it's a measure of infection. So very often people when they're analyzing CRP in relation to cardiovascular risk they delete value is greater than 10 because they they're not interested in whether or not somebody has been recently infected. CRP is also very strongly influenced

by people's medication and their anti-inflammatory medication, statins contraception and hormone replacement therapy. In this slide I show the distribution of CRP for men and women and one thing you'll notice immediately is that the distribution is highly skewed so when we're measuring when we're analyzing CRP as a dependent model or regression model for example as a dependent variable in regression model you might want to think of ways to try and normalize this distribution in order to make the assumptions underlying regression models more plausible.

So to sum up about data quality issues to keep in mind when analyzing biological datasets we need to consider the normal ranges of the biological variables if they are available. We need to be able to identify outliers and do something about the outliers. We need to hopefully I identify whether the respondent has taken any relevant medication and either control for it or maybe delete people with particular medications from the analysis if that is not central to your research question. We need to consider some of the statistical transformations for highly skewed biological dependent variables if we're looking at that in a regression modeling context. We definitely need to keep in mind the context of the blood sampling like the time of day the room temperature whether or not somebody had a recent operation if the person would have recently smoked or had food or alcohol and also keep in mind the laboratory based quality control processes in producing the biological data is it a good lab that has that has produced these biomarkers.