The final topic I want to talk about in this series of presentations on methodological considerations when doing biosocial research is about missing data and that's because missing data is a huge problem with a lot of biosocial datasets there are substantial proportions of missing biological data and there are a number of processes that can generate these missing biological data in surveys. For example there can be complete unit non-response, so somebody doesn't participate in the main interview and in the nurse visit. Or somebody can agree to participate in the main interview but they don't agree to having a nurse visit them or the nurse tries to visit them and can't get hold of the person. Or the nurse manages to visit the person but is not able to obtain a blood sample. So either the person refuses to give a blood sample or the nurse is unable to draw a blood sample. And finally the nurse might be able to get a blood sample but there are no blood analytes that are processed. So you can imagine the causes for missing biological data for each one of these steps can be quite different.

Over here I've got some of the distributions for why reasons why blood samples were obtained or not obtained from wave to of understanding society. So out of about 20,000 respondents who are eligible to take part in the nurse visit we can see that only 64% about 13,000 or so respondents actually had a blood sample taken from them. For quite a few of them, about 21% of them, either refused to give a blood sample or refused to take part in the nurse visit. So that's quite a lot of missing people but also we have people that did agree to have a nurse visit but either a blood sample could not be obtained so that's what 7% of them, or they were ineligible to give a blood sample. Maybe the nurse thought that they were too frail or they had a particular disease or health condition that meant that they shouldn't be drawing blood samples. And finally for very few people the blood samples were actually lost either in the post or lost by the lab. Now the good news is that we actually have a lot of information on people who participate in the nurse visit and people who don't participate in the nurse visit, and that comes from the rich survey information, and in addition when nurses are not able to draw a blood sample or take a particular biological data for example, they record their reasons for not being able to obtain that particular blood sample or that particular biological value. So the combination of the rich survey data as well as the nurse’s observations for why they are missing data means that this is very useful information in telling us about the reasons for missing biological samples and we can use these to explore models in which we try to take account for the differential missing biological data mechanisms. In standard, one of the standard ways of dealing with missing data is to drop a weighting procedure develop non-

response weights and there are standard variables that people consider such as you know the geographical region or what month of the year it is or how deprived the area is or the observations by the interviewer about the condition of the house, as well as the whole range of socio demographic economic indicators coming from the main interview.

So in addition to these standard sets of variables coming from the main interview or from nurses or the interviewers observations, we can also look at some of the additional information that are coming from either the nurse or from the survey data set itself that can be correlated with missing biomarkers and I'll be talking about one specific example coming from the English longitudinal study of Aging, where at wave six they measured hair analytes. So they collected samples of people's hair and measured levels of testosterone and cortisol and a whole range of biomarkers. From these hair samples as I said we get measures of hair cortisol and hair cortisone which is an integrative measure of the HPA axis, so generally higher level of indicating higher physiological stress responses. And they collected around two centimetres of hair from the back of people's head which indicates their stress levels over the previous, over the last two or three months. And you can see in this slide over here just how much two centimetres of hair is in relation to a pair of scissors. It's quite a lot of hair and in this slide you can see some of the reasons why people might be missing hair samples. So some people might not have enough hair or the hairs too thin or it might be they're taking that amount of hair from the back of their heads really makes a difference to their appearance, so they may be unwilling to give up that amount of hair. In terms of actual numbers in the ELSA wave 6 nurse visit there were about 7,400 participants, but out of them only about 2,500 have hair cortisol data, so that's a big drop in numbers and that's partly because some people were ineligible for the hair data collection they had less than two centimetres of hair, other people just flatly refused to give hair samples and again the nurses denoted, they wrote down the reasons why, and in most of the reasons was related to appearance. And furthermore funding constraints by it meant that only a subset of these hair samples could be processed to produce the hair cortisol data.

So what do we know about hair and aging? We know that baldness predominately affects men and older adults and given the importance of appearance to some participants it's quite likely that if somebody has a negative self-image of them, that makes them much more unlikely to be willing to give up some of their hair and much more likely to be missing hair cortisol data. Now if you go back to the ELSA survey that's actually a whole

range of detailed questions related to some of these missing data processes. For example there's a very good questionnaire on depressive symptoms which does cover some of the negative self-image questions so the CESD questionnaire for example, and we know that depressive symptoms and stress levels are interlinked, so it might well be that the people who are missing hair cortisol data are, some of them are missing, because they have a negative self-image and they're more likely to be depressed. So if we just analyze the complete cases people on whom we just have a hair cortisol data and depressive symptoms, we might get sort of a biased estimate off of a number of associations.

In this graph on the left hand side we're looking at predictors of who has given, of who has hair cortisol data and in the left hand side we have an age and gender interaction, in the blue bars are our men and brown bars are women, and we can see that for men in particularly men aged 60-65, 70-75 they're much less likely to have hair cortisol data compared to women of the same age groups. And on the right hand side we have depressive symptoms as a predictor of having hair cortisol data and for those who have who score four or more on the CESD questionnaire there, in other words they have high levels of depressive symptoms they're much less likely to have hair cortisol data.

So we can use these kinds of information to develop weights or non-response weights for missing hair cortisol data and what happens when we when we take account of these kinds of differential patterns of non-response. Well on the left-hand side is the complete case analysis showing the association between depressive symptoms and levels of cortisol derived from hair, and we can see that there is an association, so people who have higher depressants more depressive symptoms have higher level of hair cortisol. That's not surprising. What is perhaps more informative is on the right hand side which it does take into account this differential pattern of missing hair cortisol data, taking account that the fact that the people who are missing hair cortisol some of the people who are missing hair cortisol they are also scoring higher on depressive symptoms and we have a stronger association between depressive symptoms and levels of cortisol. So just to remind us it's the rich survey and nurse observation data that allows us to discover some of these factors that are correlated both with the missingness mechanism as well as our outcome of interests, and if we make inference based on, just on complete case analysis that may be bias, if we don't take care of such factors. And so the good news, as I said, is that we do have all these rich sources of data, both missing biological data coming from the nurse observations, as well as rich sources of data

coming from the survey observations to explore reasons for why these data are missing which we can then take into account in a number of statistical ways.

So I just like to recap the complete series of talks today about some of the methodological considerations to keep in mind when doing bio social research which is about the need for having a bio social research framework, the need for keeping in mind some of the data quality issues and need to keep in mind some of the missing data procedures and processes that happen in bio social research. And a biosocial theoretical framework is absolutely key because otherwise we just have a whole range of correlations without that is meaningless and often unscientific and not reproducible, and that's why we really need to have good interdisciplinary and multidisciplinary research teams to investigate these associations between social sciences and biological data. In terms of data quality processes, we need to consider the normal ranges of the biological variables to identify the relevant outliers to take into account things that can affect the biomarkers, like the medication use, the sampling, the time of day of the blood sample, the room temperature for example. We also need to keep in mind quality control processes, the QC processes in transformations if we're looking at dependent biological variables. And finally we need to keep, to identify the relevant predictors of missing biological data to use in non-response methods, and that's coming from both the rich social attitudinal data from the surveys as well as the nurse’s observations.