Good evening ladies and gentlemen and it's a real pleasure to welcome you here this evening because really without your contribution to the study none of what I'm going to give you an overview of tonight would possible, Andrew has already given a flavour of just how powerful this resource is becoming and I hope to convey that tonight in a bit more detail.

The overall strategy why do we set this up what was the purpose of it? You are amongst half a million men and women who were recruited into UK Biobank starting in about 2007 and the only selection criteria for inviting you was that you were aged between 40 and 69 at the time of your invitation. As you will recall we asked you a lot of questions, baseline, I use this term baseline, that just means your first visit that's the sort of baseline that we set you at and you will hear that term quite frequently. We took physical measures and we stored blood, and from later participants we also stored samples of saliva and urine and they've been stored in such a way that we don't try to predict how technology will develop their stored in a way that will allow as many analyses in the future as possible, but absolutely critical, and is one of the things that makes this study unique in the UK and internationally is that you have given us your consent that we can follow up your electronic medical records and the UK's almost uniquely positions in that regard and therefore to do these kinds of studies.

This study is described as a prospective, in other words you weren't recruited on the basis of a pre-existing disease or condition there is disease within the population but that was not the reason for recruiting people and it allows us to follow your health over time and assess the full effects of the particular exposure, for example smoking, not just on the single outcome for example lung cancer, but on all health outcomes, that maybe cardiovascular disease it may be colorectal cancer, it may be all sorts of things. Normally with these studies, the UK has a reputation, a well-earned reputation in this area of doing large-scale population studies but the studies tend to be large which meant that the amount of data that you could collect on any individual would be relatively small, or they would be small in size and very rich in data, but because of the way these studies have now been set up and the ability to use technology we can get the data depth and the data breadth on individuals, but really importantly why

500,000? Well because it's prospective because we don't invite people because they have a particular condition within that 500,000 sufficiently large numbers of people will develop these different conditions and that will allow us to untangle and unpick the quite, the delicate interplay of these different causes that lead to these various outcomes. This just shows you a prediction of the number of what are called incident disease outcomes that we expect in this study. So again just some definition of terms we talk about prevalent disease which is disease that might have been present in people when they came to the assessment and incident disease is disease which occurs after the recruitment and you can see within a population of 500,000 people we expect a very wide range of diseases in quite substantial numbers and it's these numbers that give you the power to start to unpack these really quite subtle effects, so for example you can see MI is heart attack so we will have, we have had several cases of this already and you can see by 2022 we'll have 28,000. Now people ask us at these presentations are we as participants, are we typical of the UK population. Well you are not typical in one regard you are about twice as healthy and these numbers were put together these estimates were put together rather conservatively recognizing that so-called healthy responder effect.

Just to emphasize the size though, this figure here shows what's called a meta-analysis, and a meta-analysis is simply taking lots of studies and combining the data to increase the power. This is looking at the risk the relative risk on the on the y axis down here for two known risk factors, along the bottom here the x-axis that's your blood pressure the point at which the heart pushes the blood out of the heart. The so called systolic blood pressure, and on this side here it's grouped by age, the participants in these studies are clustered into age and this was done on this so called the prospective studies collaboration which was run out of Oxford. This took a million people, if you take 5000 of those people and you look at the relative risk of heart disease, of death from heart disease you can see as you go in increasing blood pressure there is a general left-to-right trend it appears that as blood pressure increases your risk of heart attack increases. These lines here you may just be able to see in these lines individual points here, and these represent the actual data point and these lines represent if you like the measure of uncertainty so the data falls somewhere within those lines and what that means is it's incredibly noisy, it's very hard to draw any firm conclusions about trends and associations of risk with blood pressure and with age. If you go to looking at not 5,000 people but 50,000 people you can start to see some trends emerging these so-called error bars, these lines are

fore-shortened and so the degree of confidence is much higher albeit still quite large and you can see general trends from left to right with blood pressure and also possibly with age, there are some rather unusual features to these curves though, you can see at the highest age range from 80 to 89 it appears to flatten out and possibly even drop with higher blood pressure.

At the lower age range down here it appears hockey stick shaped, so that lower blood pressure appears as to slightly increase your risk of heart disease, so is that real? But when you go to half a million people it really comes into very, very sharp resolution and you can see this very tight correlation between increasing blood pressure and increasing age and the risk of heart attack it is very, very, very, very clear it's like going from using the naked eye, to a magnifying glass; to a microscope it gives you incredible resolution. This is how we recruited the half a million people, you were recruited at one of these centres around here and you can see that there are other regional centres this is relevant to the work that Steve is going to talk about on imaging. There are about 26 of these dotted around Great Britain. I mentioned that you were typically more healthy than the British population; these are the characteristics of our study population. Slightly more women than men, our age range just over half in the 40 to 59 ears at the time of recruitment not now obviously and 43% in the older age range.

All strata now what does that mean? All strata means social democratic strata in terms of education, in terms of sociodemographic status, in terms of disposable income, scores of deprivation are represented about 85 percent of our participants come from urban addresses and about 95 percent of the population described themselves as white, about five and a half percent as other ethnicities, so it's a pretty good picture of the population. This will be familiar to you this is the baseline again I use this term the baseline assessment visit; these light blue boxes are the stations that most of you will have done. So you did the touchscreens station when you gave your consent and then asked all these different questions which you answered on the touch screen format, we did an interview with you and took two measures of blood pressure, blood pressure tends to be quite variable she took a measure at the start of the interview and a measure at the end. And then these fairly standard, but very informative physical measures. Hand grip you may have heard about this on the radio just recently, they are starting to show correlations with heart disease it’s a very good predictor of heart disease.

Your height standing and sitting. The ratio, we could derive this, but the measurement around your waist and you hip, your weight and your an estimate of your body fat by passing a small current through you which you probably wouldn't have even felt and assessment of your lungs and an assessment of your bone density by measuring the properties of ultrasound through your heel and then you gave samples of blood and urine and that was that was that for most of you. As this bedded in and started to work very well, we received additional funding to add some measures to this, so we added some additional measures of cognitive function and a hearing test we implemented the Royal National Institute for the deaf, speech in noise test, where you have three numbers against the background noise and the level of the background noise gradually increases to the point where you can't distinguish the numbers and that's your hearing threshold and it worked very well. We looked at a measure of the stiffness of the of the vessels in the finger just using a finger clip, it's called vascular activity how flexible the vasculature is. Some very extensive measures of the eye, everything from a fairly simple so called logMAR chart where you see that the letters on the wall and they gradually get smaller, right the way through to the so-called 3d optical coherence which is an extraordinary technique which allows you to build up a three-dimensional picture of the retina at the back of the eye, so you can take in the computer slices through the layers in the back of the eye, it's an extraordinary technique. For some people we had an assessment of physical fitness on the static bicycle with an assessment of heart rate and then we took some additional samples so that's really the baseline resource. Many of you will recognize these, these are the vessels into which you gave your blood I mention that we were collecting a range of samples and they were stored in such a way so as not to preclude future analysis and the value of that has already started to become clear. Exactly how these are put together actually isn't the purpose of this slide the point of this is to show you a few things, as I say there are a variety of tubes here which make these tubes useful for different things, but this really was to show you all of these tubes are are bar-coded ok and they were linked with your participant identifying number when these tubes were collected and effectively this gave a very secure and very robust link between you as an individual and the samples and everything that happens downstream to them. So this is the first point where we actually break the link in the processing centre with your ID. Obviously the link is held securely by the principal investigator, but when these came to us in the centre in Cheadle we couldn't tell you which belonged to which person, so it is very secure. These are 12 digits, they have a checksum in

them and the error rate in reading these is about one in a billion and it's really important because this is the starting point of a quite complex downstream process where everything that happens to your samples is then linked back through this tree back up to your participant identifying number so it has to be very very robust. This is the process that you were involved in you were one of seven hundred participants a day, we would typically operate six assessment centres dotted around the country at any one time and recruit around about a hundred and twenty people a day. That would produce about 5,000 of the vacutainers, these vessels that your samples were collected into and then they would be aliquoted or split up into their constituent parts to produce 21,000 samples per day, so you can see the point was making about the barcodes the idea of trying to track this manually with paper and pencil you would just lose track of it in half a day, it would simply not be possible, and at the end of the study we have forty million aliquots, so keeping track of that in a manageable way this really is the only affordable and feasible way of doing it. This is where the samples are stored this is for those of you who know this is just in South Stockport opposite Cheadle Heath police station and this is actually where your samples are stored.

We've split your samples between two sites to get the ultimate protection, if we have an absolute disaster, if a plane lands on this site we still have a copy of your samples stored in Wythenshawe. This is referred to as the working archive this is a fully automated, minus eighty store, and I'll expand on that a little bit in a moment, but about 10 million samples are stored in that store there. If you look inside this, this is what you would see. There are 9 independent storage towers which store ten million samples at about minus eighty degrees C and they're stored behind the sort of insulating bricks on trays. All of the work is done by this workhorse robot this yellow robot that you can see in the middle and this moves up and down this aisle at great speed and either puts the samples away or retrieves them it can do this in complete security, your samples never thaw when it does all the manipulation and it can do it very quickly but that's not really the point, the real value that this brings is it will pick the samples with complete accuracy, if you were doing this as a human being you would have great big gloves on it would be very demanding, it's very fiddly and it's very boring. We know from testing this that you can get up to a 10 percent error rate. Now you're looking for associations with diseases that actually have quite small effects and if you start to pick the wrong samples what you're effectively doing is creating false positives and that really really causes, has an enormously detrimental effect on the power of your study so itis true to say that the speed

and capability at this thing is important but it's the accuracy to the quality of the science which is more important. This is a backup site, sunny day in wythenshawe and we store our backup samples about five million samples in these very large insulated tanks here; there's about five million of them. These are stored at about minus 196, so these are stored in nitrogen vapour and that really will keep anything in good condition. It's a warehouse, it's fit for purpose, no frills but you can see around the edge the kind of attention, this is a priceless resource, so we have implemented things like a fire suppression system because we have very sensitive detection but our concern would be if a fire starts obviously the fire brigade would turn up the fire may have got hold and then they would spray things with foam and so forth, they would do their job. This would put the fire out before that happens. Okay so that's a bit of a whistle stop tour to how we set the study up. What was done and how we got to what we think of as the core resource or the baseline resource.

What are we actually doing now, well critically having spent all this time and money and you having spent all your time giving all these samples and data and so forth the resource is now being used and you will hear some research from two of our collaborators this evening, so we are providing access to the resource. Critically we are starting to follow people's health through the electronic medical records the study is extremely powerful anyway but once you can start to follow people's health it becomes really very very valuable indeed because you can start to see you know what's happening to people. Then what has actually happened is this approach to these large-scale studies has proved very successful and it's attracted additional funding so we're doing a lot of what we think of as enhancements to the study that weren't envisaged when the study was first conceived in 2001, 2002. Some of you may have received requests for questionnaires over the internet looking at cognitive function or diet, there are more to come. Some of you may have received a wrist worn device that measures, it gives a very accurate quantitative assessment of your physical activity and we're doing that in a hundred thousand participants. We're measuring biochemical markers in the blood and urine of all five hundred thousand people that are informative about disease, but the two that I'm going to concentrate on tonight the last one will lead into Steve's talk is Genotyping of all 500,000 people, and I'll explain what that means in a moment and then the imaging of a hundred thousand people, really just completely ground-breaking work.

Coming back to the providing access to the resource it is really starting to attract a great deal of attention and this slide really shows you the range of diseases and conditions that you can look at. You remember in one of my opening remarks I said you can look at a whole range of conditions you not just focusing on one particular condition, and along the bottom here this may be a test of visual acuity you can see the range of diseases that this study is being used for. It is quite remarkable really and these numbers here along just above the bars show the number of applications that we've had from researchers to work in these areas and this this really is really just a foretaste of things to come because it's more and more of these data come out and the real power of the studies that we have, lots and lots of data and different types of data on the same individuals these numbers are going to grow very, very quickly. We're following up people as I said you know to do anything on half a million people we obviously want something which is comprehensive we don't want large gaps in our understanding of people's health but it has to be scalable to half a million, it's fine to do this for five hundred people, but it has to be scalable to half a million and of course affordable. For us as I mentioned we are probably unique internationally, all of our participants are registered with the National Health general practitioner that's how we got your contact details and of course the NHS provides the great majority of healthcare so we have a very, very good picture of the majority of healthcare which is provided to our participants and of course we have national data sets about health care and what happens to people. We have a register of deaths and the causes and we have links to that, there are cancer registrations and there are several databases that provide more detail, if you go into hospital a lot of information is recorded about you when you go in, everything that happens to you is recorded and we can have access to those data and of course primary care data or your general practitioner data is also available, that's a bit more tricky to link to it's a rather fragmented landscape so we are in the process of linking to this in England. I mentioned the web based tools, again they are affordable fits it's very simple to send out an email for people to click on a link and fill them in Of course you can do repeat administrations and that's important for things like diet because your diet will change over the course of the year. It's a very efficient way of collecting and coding data if you do a traditional paper question you have to have somebody has to go through and code all the answers into a computer this person responded yes to this no to that and so on, this of course is done automatically because it is all done electronically and of course the analysis that follows on from that. I mentioned the dietary questionnaire and the

great advantage of this is you could assess conditions that are quite hard to find out about from health records, for example things like cognitive function the real power of this is if you're interested in Alzheimer's disease you can learn a lot about the disease at the time that people present but of course by then much of the damage may have been done. So you may not be able to identify the causative factors, if you can start to see characteristics of people and correlate that with functional decline, cognitive function decline, that then go on to develop Alzheimer's then that might lead you down a different research avenues. Then things like mental health and quality of life.

Ok so on to the genotyping and the imaging. A quick crammer on genotyping I said I'd explain this a little bit, so you probably hear on the news things about genome and genomics and DNA and so forth. The human genome, in other words in any one of you your entire DNA compliment, all of your genes contained on your chromosomes is comprised of about three billion bases, chemical letters very similar chemical letters that make up the DNA helix. These are carried on the chromosomes as you will know. Each human chromosome carries a unique set of genes which make specific proteins which make us what we are. Each of us has about 30,000 genes which is only 2 percent of the DNA, so a lot of the DNA doesn't actually finish up, the information in the DNA doesn't actually finish up in the protein, so there's a lot of other functions of the DNA, which aren't necessarily involved directly in the proteins but this genetic diversity together with environmental influence is what makes human beings, each human being different from each other and within this room most of us are unrelated we are 99.9% identical to each other at the DNA sequence level and in fact we're actually 99% similar to this fellow, this rather thoughtful looking chap here, we are we are 99% identical at the DNA sequence level to him. But it's the variation and it's only one in a thousand chemical letters that doesn't seem very much but it's that variation that causes a difference between us and drives evolutionary change. Now the most common, there are lots of different types of variation in the genome, there are many different types but the most common form is a thing called a single nucleotide polymorphisms, a snip ok. What does that mean? If you look at the sequence of two very, very similar people at their DNA. This is person A this is person B, on the bottom you can see their sequence of this particular region of their DNA is very very similar however there is a difference at a single chemical letter or single nucleotide as it is referred to, hence the name single nucleotide polymorphism there you have a C and here you have a T. So

that would be a SNP or a single nucleotide polymorphism. Ok so what we are doing at UK Biobank is really on a very large scale indeed. High-quality genomic DNA, in other words the DNA from within your normal cells in this case the white cells that we extracted from your blood is extracted on automated systems in our centre in Cheadle and you can see the systems here. The DNS is genotyped, in other words these points of variation across the genome are measured by our partners in Affymetrix and then the genotypes are called, in other words there are signs so using the data they will say at this particular position the chemical letter is a C or G or A or whatever it happens to be. Then they are quality checks with our collaborators in Oxford. These data though, so we will measure eight hundred and twenty thousand points of variation in your DNA but the data are enormously enhanced through imputation and I'll explain a little bit about that in just a moment.

Here's an example that that really broke ground in in this entire genotyping area and this is the Wellcome Trust case control consortium, what they did was they took two thousand cases each of seven very well characterized diseases and you can see them here, so we've got coronary artery disease, bipolar disorder sometimes known as manic depression, we have Crohn's disease, hypertension and so on and what they did was they used a standard set of controls and they measured about four hundred and fifty thousand points of variation across these peoples DNA. They couldn't do the level of imputation that we can do that I'll come onto in a moment if you focused on this, if you just look at coronary artery disease this thing here is referred to as a Manhattan plot and each of these blue blocks represent the likelihood of those markers being associated with the disease on each chromosome. So in chromosome 1 you can see these different markers here that may or may not be associated with disease actually they're not because they're down in the noise, the statistical probability of those being associated with disease is too low to confidently assign them. However you can see here in all these individuals as a marker pops up here and the chances of those occurring purely by chance is one in 10 to the 15 or a thousand billion to one so that looks like a very good candidate for further investigation. Ok I just need to come back to this point about imputation because this is incredibly powerful and it shows the rate at which this error is moving on and imputation allows you to infer the presence of other sequences in the DNA without actually measuring them. So imagine you had a dictionary ok but you could only read every thousandth entry in the dictionary, so this is gene this is the dictionary that you have and you

can only read every thousand entry but imagine if you had reference dictionaries where you did have all the entries, what that would allow you to do with real confidence using these two dictionaries is to predict with a high degree of certainty the entries on either side of it, so you could start to fill in, in your dictionary what's either side of it, and of course the further away you get from the measured entry in this case G the less certain you become, it becomes a bit cloudy. You can do this of course with DNA, by comparing it to reference sequences which have been fully sequenced you can assign markers either side of your measured markers. Now I said this enormously enhances the data, we have measured about 800 to 820,000 markers in your DNA, using these approaches you can infer or impute the presence of another 73 million markers with a high degree of confidence and by getting up to those sorts of numbers you are actually measuring or calling about one and a half percent of the Genome.

It really is an extraordinary resource and that's going to increase as we get more and better reference sequences. I'm going to finish just on this now but it's a tee up to Steve's talk the other really exciting area is this large-scale imaging project that we're doing. Why are we doing such a large-scale imaging project? Well of course one of the things about you is you are already very well characterized and we have your permission to link to your health records, so we can link these data onto your health records. Of course we increase the power of the study not only through size but we have measurement precision. Some of the measurements you can take with these techniques are incredibly precise and what makes this perhaps even more valuable, is many studies exist which are in the kind of 1,000 2,000 5,000 people but they tend to have only a single measurement modality. We will have multimodal imaging as you will see, so we won't have a single measurement on you we will have lots of measurements on you. It will be the largest study of its kind in the world. I'm not going to steal Steve's thunder but some of you may have done this already, this is the imaging facility that we set up in our warehouse in Cheadle these are the MR machines, we have two of these. That carries out measures on the brain, on the heart and on the abdomen. We have a machine here that measures a low-energy x-ray analysis of the bones and joints and it will also give you an estimate of body fat and then a device here which looks at the carotid artery where the carotid artery splits in two in the neck it looks at characteristics of the carotid artery. So we'll have these five measures hopefully on a hundred thousand people.

So there isn't really time to go into all of these different measures so rather than give little vignettes of these measures I thought I would just demonstrate how valuable these data are in one area. This is work done by Tony Goldstone and Jimmy Bell at Hammer Smith at Imperial College and they're particularly interested in the distribution of fat in the abdomen and the impact that has on health. An accurate measure of fat distribution and looking at the link to risk and to health outcome. Of course it is more informative than body mass index this kind of general description of how big you are and whether you are obese. This term here WHR, is waist hip ratio which is also quite an informative measure albeit somewhat low resolution. They of course are interested in the volume but also the distribution of fat as you'll see and its link to metabolic diseases like diabetes and heart disease and so on. They're very interested in the distribution of fat in the liver and in the muscle as predictors and outcomes of disease and the interactions between what they refer to as body phenotyping of your body type and I'll expand on this in a moment and things like genetics and lifestyle factors. If you can start to link these data to all of those Genotyping data that I spoke about you can see the power of this. There are no large in-depth studies to date, as I say, a 1000, 1500 people but it also, there are some really nice things they can do with this looking at the impact of physical activity, we're measuring physical I described and also they are also interesting in things like the impact of calorific reduction. Just to give you an example these are beautiful images produced by Jimmy and Tony and they are sections, transverse sections, imagine a slice through here of 9 individuals, these are nine separate individuals and they are chosen because their body mass index is pretty much the same they're all and so on, that fits very neatly in the normal healthy range when you see these kind of rankings of body mass index. What you also see though is the very different distribution of fat. TAT is total adipose, it's the amount of fat in them essentially you can see in this individual here there is 13 litres of body fat with a BMI of 23.5 this individual has a BMI at the bottom right down here of 23.6 but has twice the amount of fat and you can see in this individual here most of it is just under the skin so-called subcutaneous, this is the belly button here, the umbilicus. This person there is a lot under the skin but there's also a lot around the organs here as well you can see this all round the organs in this individual here. This individual here has a BMI higher than this person here 24 versus 23 and yet very low levels of fat. BMI of 24 but only 12 litres of fat in their body. So it's a much more precise measure What they started to realize was that if you take the descriptions of individuals these body phenotypes these body types if you like as lean, overweight, obese and

morbidly obese these are definitions you would have heard about from BMI and actually look at the so called intra-abdominal fat, the total fat if you like you can see there's a huge degree of overlap people who were described as morbidly obese on BMI actually have lower intra-abdominal fat than some people that are in the lean. That doesn't mean that BMI is not a valuable measure. It just means that these are more precise measures. So Tony and Jimmie coin these Terms Tofis, thin on the outside fat on the inside, you can see here so they are thin on the outside but fat on the inside. Of course there is the reverse of that which is Foti which is fat outside thin inside, which is really interesting. The influence of dieting, this one always generates some discussion so there are three individuals along here these are three separate individuals and these are the same individuals underneath after a weight loss of 33 kilograms which is quite a lot of weight to lose but actually when you look at it these people who are dieting you can see that the weight goes is from around the so-called subcutaneous, the stuff underneath the skin you can see all these individuals they lost a lot of weight around here. A lot of the weight has gone from around there around the buttocks here but there's still a lot of the fat around the organs and that's the stuff which is associated with metabolic disease. So dieting alone in this particular regard needs to be accompanied by exercise. There isn't time to show you the equivalent where if you look at the impact of exercise you can see that exercise will shift the visceral fat, the fat around the organs.

So in summary I hope I've got across to what a globally unique resource you are involved in you've been responsible for setting up and how it's now starting to become widely used and it will be used to two completely different level of detail to understand the causes of common disease. Of course as more and more data are added whether it's from our own endeavours or from researchers using it, because it's one of the few stipulations of people using this that the data returned to the resource to enrich it. Of course it becomes increasingly valuable more and more data are available. New technologies are offering the potential for really ground-breaking discoveries. I've been with Biobank 11 years and was involved in that the very first endeavours to set up the sample collection protocol and when we did that in 2003, if we had sat there and said we will be able to measure eight hundred and twenty thousand points of variation in the DNA and impute or infer another 73 million for 32 pounds per participant you would not have been taken seriously and in 10 years that's how far we've come. The first human genome sequence was published in the year 2000 it cost about a two billion dollars

and it took three years to do, but you can now doing it in a day for a thousand pounds. It is absolutely incredible the rate of technology change, of course that creates new challenges downstream. The way this study has been delivered and continues to be delivered

I think it's been recognized nationally and actually internationally too and I think, I genuinely believe this, you know it is something which the northwest and the UK can be proud. I thank you very much for coming tonight. Thank you for your time.