Thank you very much. It's a pleasure to be here, as always, and talking about just the amazing stuff that's happening here in UK Biobank. Let's see, all right, great, so to start with we've talked a lot about genome-wide association studies and sequencing. We've talked some about phenome-wide association studies as well. That's going to be the focus of my talk and just to orient us, essentially what we're doing is thinking about it and looking at variables and exploring what phenotypes and the range of phenotypes that are available and associated with that. Really anchoring on the fact that we have richly systematically phenotyped sets of individuals such as the UK Biobank and other electronic health record datasets, which is where this started. Usually that's based on things like billing codes, but I don't want to limit us there; you can think about laboratory values, you could think about natural language processing and things like that as well.

Most of it's been based on billing codes, so to start with, I want to orient us to a discovery study out of the electronic medical records and genomics networking merge in the US. It was five sites that worked together for a carefully-validated phenotype that we used codes, labs, medications, natural language processing to find these cases and manually validate who was a case for presumptive autoimmune hypothyroidism and the control. We identified a thyroid transcription factor that was associated and replicated thus. Then we took the same variant that was found here and did a PheWAS on that variant in a slightly larger population that was unselected or much larger population, unselected for any given phenotype, and hypothyroidism was the highest associated phenotype there. But we also had some other thyroid diseases that came up and things like atrial flutter as associated, which we know that there's hypothyroid or less likely to manifest with atrial flutter.

This gives opportunity to look at the performance of these two methods, and so on the left we have the schematic to the algorithm we use and then we use these mappings, called PheWAS codes or Phe codes. We usually say there have to be two or more that map into that Phe code. You can see the odds ratios are essentially the same between the two approaches. Within our population of individuals with any merge we identified more cases with the PheWAS codes than we did with the algorithm. So there's many approaches of PheWAS and just talking about that, in most used billing codes in the US, it's been historically ICD-9, with the clinical modifications, and now ICD-10 after 2015. There's about 65,000 ICD-10 CM codes and on the right you can see some of the ways this works. The Phe codes have numbers, kind of like with ICD-9 codes but they're actually not. Then what we do is we group like codes now across ICD-9 and ICD-10 and ICD-10 CM codes to a given Phe code.

All the type 1 codes come together, which is not obvious from the ICD-9 coding group system, and then each of those also define ranges of control groups. In addition to the Phe code groupings, there are other groupings in the US. There's the AHRQ has released some software that groups things into about 300 diseases. is another thing. You can also use raw ICD codes for instance, and that gives you a challenge across of course mapping between ICD-9 and ICD-10. You can do many other things. Survey data has been run across the UK Biobank and other things that I've talked about, like the procedures. So here's another example of PheWAS driven by EHR data, looking at imputed HLA types into the two and four-digit types of HLA.
You can see quickly are the highlights; the fact that there are different associations between class 1 and class 2 HLA alleles and what helps you think about the range of associations.

Overall I think there was 100 or so significant associations, most of which were known, and a few new ones but what's more interesting, by doing it in a single population you can actually look across those phenotypes and then look for pleiotropy and see if you adjust a condition in one or the other, do you see that they're truly independent associations? You can also see in a given HLA type that two HLA types that may put you similarly at risk of rheumatoid arthritis may have differential effect for your risk on type 1 diabetes, for instance. So that is a tool that you can rapidly explore using this kind of technique.

An important aspect is validating its efficacy, so one of the early things we did using our ICD-9 codes across eMERGE was replicating known associations in the GWAS catalogue. We found 86 phenotypes that could be represented in an electronic health record, and a number of SNIPs, about 750 overall SNIP phenotype pairs. Overall we replicated 210 of them across a number of disease classifications and 66 per cent of those for which we were adequately powered in this population of 13,000 people as well as identifying some novel associations, the top of which we replicated.

It also allows us to actually compare the effect size, so here you see something that you would expect to see in that the effect sizes from the PheWAS studies are typically a little bit lower than what's the GWAS catalogue. Now, some of that's probably due to the but some of it's also due to the phenotype being not quite as accurate. It helps you think about the ones where you have the most error, so the most common error was really universally type 1 diabetes is often miscoded. In fact, 96 per cent of the time, we found type 1 diabetics had type 2 diagnosis codes. So it made it - and the reverse is true 56 per cent of the time, so it caused a lot of inaccuracy in the type 1 diabetes phenotype and we had trouble replicating some of those SNIPs. We've actually instituted methods to fix that problem and we can recover those associations.

Here's a way you can use PheWAS in concert with a GWAS. We did a GWAS at eMERGE, looking at longitudinal risk of cardiovascular disease on a statin, and found variants that are tied to expression of lipoprotein A were associated with that outcome as a longitudinal analysis. That risk is increased for those that have ideal cholesterol levels at less than 70, so we looked at a PheWAS of this locus. As you'd expect, you see coronary atherosclerosis near the top, and fortunately we see most of the phenotypes are ones we would expect to see, which it gets the question of: if you were to target this with a medication, what potential effects would you see? One of the things that's interesting and wouldn't have been on our radar screen is this point over here, which is not quite statistically significant, was lung cancer. This is a relatively small population of 13,000 people. As it's explored more, maybe that will turn out to be true or not, but it is a rapid tool for highlighting, especially when you think about the scale of UK Biobank.

I mentioned mapping these two ICD-10 and ICD-10 CM codes; it just shows a little bit of a process and the vocabulary and systems that we use in the process, with some manual validation. It is in what we call still in beta form but you can see it covers about 90 per cent of the billed ICD codes in the UK Biobank. Amongst the ten per cent that aren't there, most of those are not actually disease codes; only a small fraction of
those represent true disease codes. We did an evaluation using our data, with ICD-9 and ICD-10 codes in terms of PheWAS and you can see that the effect sizes in this population was essentially the same for these two known associations with that SNIP.

I want to give a few examples, actually shared this earlier: doing a PheWAS in UK Biobank and just tons of association associated with atrial fibrillation, genetic risk score for a-fib. When they conditioned for the cardiovascular phenotypes, essentially, those associations went away. But it shows the power of a huge population, it shows lots of things you would expect to see. Here's another one for systolic blood pressure on a large GWAS that was done across Million Veteran Program, as well as the UK Biobank. Just a number of associations showing up with systolic blood pressure. They also did it with diastolic blood pressure and pulse pressure to show that some of these phenotypes overlap. You see phenotypes that are not exactly associated with cardiovascular disease in here coming out as well, endocrine being one of the more common ones.

Here's a resource, Kristen also talked about the SAGE approach to use saddlepoint approximation to create an efficient and accurate way of calculating these kind of results at scale for the UK Biobank. They have produced a website where you can explore those phenotypes, calculated using the same approaches for Phe codes across UK Biobank. This just shows a particular a-fib SNIP in that website, and the URL is there at the bottom. We talked about this and looking at individual phenotypes. I want to spend the last few minutes talking about phenotypes in clusters and how we think of them. If you think about Mendelian disease is a classic example that are often syndromic, presenting with many different features. Those features are what we may bill in the electronic medical record as physicians. But it doesn't necessarily represent… The disease is not always recognised, or maybe recognised later in the disease course as we heard about earlier, with haemochromatosis.

Through the Online Mendelian Inheritance in Man resource and the link to Human Phenotype Ontology, we can go from a Mendelian disease to a list of features of that disease, which have a vocabulary behind them. Our lab mapped those HPO features to Phe codes, so basically allowing you to translate features into HER phenotypes and then similar to a polygenic risk score, creating a phenotype risk score that looks similar in process, so aggregating phenotypes by their weights to produce a score for an individual. Essentially you can crank this out across anything for which you have a map and do it at scale. So let's look at cystic fibrosis, a number of features from OMEN and each of those is mapped to a human phenotype ontology code. So I'm using our Phe code ontology of around 1800 phenotypes. You can map the ones that line up fairly well to the . They're not all exact matches; some are better matches than others, and then some that we don't have in the HR which we're familiar with.

Let's play that out in a couple of individuals, hypothetical different conditions. I mentioned they're weighted so features like bronchiectasis have a higher weight than features like asthma. When you go across this, individuals get a different score and what you find is you can separate cases and controls for cystic fibrosis, just using the features of disease. We're not using the disease label but in this example we use mainly validated cases versus controls, who don't have any evidence of the disease in the text record. We see a very
significant result and we've actually done this for 15 other diseases now and in every case, except for one, we've seen very strong separation between cases and controls. The one exception is phenylketonuria, which as you know in the US is on essentially every newborn screening test and if you avoid phenylalanine exposure, you don't actually see the manifestations of the disease. It gives you a test of the effectiveness in newborn screening and removing the future disease in the population because they generally do not have elevated scores.

We turned this on a population of 21,000 people that had exome array genotyping, and looked at 6000 variants that were rare at one per cent level or less. We found 18 significant associations, most of which were novel, and importantly we were able to change the ACG clinical interpretations for each of these variants towards a likely pathogenic or pathogenic. So this using our population is a paradigm that I think can be explored with a larger rich phenotype population such as what is in the UK Biobank. I want to end with recognition of some of the main people contributing to this work. The middle row is probably the most important row, as the folks actually doing the work. Thank you very much.