Sarah Hiddleston: Hello and welcome to this Nature Research Custom Media broadcast titled Proteomics for Precision Neuroscience: The Power of Protein Analysis. My name is Sarah Hiddleston, and I will be your moderator. Today's webcast is sponsored by SomaLogic.

Alzheimer's Disease and other neurodegenerative disorders involve a complex interplay of genetic, molecular, and environmental underpinnings. This webcast will highlight how academic, industry, and government researchers are currently measuring protein abundance and function via multiplex proteomics to build more detailed characterizations of the biological systems underlying neurogenerative diseases. Combining large-scale proteomic studies with advanced bioinformatics can facilitate the discovery of new protein markers for diagnostic purposes, disease progression markers, molecular targets for drug discovery, disease subpopulations, and risk prediction algorithms.

We'll begin with presentations from three scientists working in three different areas, who are all with us in the studio today. They are Dr. Christopher Whelan, director of Data Science at Janssen Research and Development, LLC, who trained in neuroscience and statistical genetics before moving into industry. We also have Dr. Carlos Cruchaga, director NeuroGenomics and Informatics Center at Washington University School of Medicine in St. Louis. He is a human genomicist with expertise in multiomics, informatics, and neurodegeneration.

Finally, we have Dr. Keenan Walker. He is chief of the Multimodal Imaging of Neurodegenerative Disease or MIND Unit at the National Institute on Aging. His current research program focuses on understanding the role of abnormal immune function in Alzheimer's disease and late-life cognitive decline.

We will then move on to a question and answer session with you, the audience. You can ask a question at any point you wish throughout the webcast. To do so, please type your question in where it says, "Type your questions here," and then press "Submit," and we will answer them at the end of the session today. Now over to our speakers.

Dr. Christopher Whelan: Thank you, Sarah. There couldn't be a better time to speak about proteomics with enormous excitement around this field just in the last 24 hours, but I'm going to begin by speaking about another field, genomics. It's been roughly 20 years, give or take, since the completion of the Human Genome Project, and that was heralded, rightly so, as the first great technological triumph of the 21st century. Figures like Bill Clinton, Tony Blair, Francis Collins, they all predicted that within the next 20 years, we would have cures for Alzheimer's, Parkinson's, cancer, diabetes due to genomics. In many ways, there was a promise of precision medicine.

Now, like many things, the reality of the ensuing 20 years looked a little bit different to the expectations. After the completion of the Human Genome Project, the field entered an error that you could call the "candidate gene" era, in which a handful of gene variants would be tested in maybe 50 cases and 50 controls. By and large, most of those hypothesis-driven studies were not replicated. Of course, there are exceptions like the discovery of APOE, but by and large, this was not a particularly fruitful era.

Then there was a course correction in the mid-2000s where we entered the genome-wide association study era in which we perform hypothesis-free screens across the genome from chromosomes 1 to 23, and the sample size has increased starting with a few thousand cases and controls moving into tens of thousands of cases and controls. Then at the population scale, hundreds of thousands and in some cases millions of participants. We've seen some success in this population-scale GWAS era that I'd actually argued that we're entering a new era, the post-GWAS era. Depending on who you ask we're either very close to precision medicine or we're still a long way away when it comes to genetics.

The skepticism around the ability to transform genetic discoveries into precision medicine is relatively well-founded. When we think about what precision medicine means, ultimately, it's about finding the right drug for the right patient at the right time. In the context of genomics, the right drug, most of our drugs don't target genes. Of course, we have gene therapies and other approaches, but the most widely applied kinds of approaches are not targeting genes. For the right patient, of course, genetics is incredibly useful for rare diseases with high-impact mutations, but for more commonly occurring illnesses, we still need to wait and see about the clinical utility of polygenic risk scores.

Then finally, the right time, genes are static. It can be tricky to actually put a temporal component using genetics. Now, this stands in stark contrast to proteomics. With proteomics, we can find the right drug. Most of our drugs target proteins. We can find the right patients because there are already many clinically validated and clinically applied tools that are based on blood proteins like ApoB for hypercholesterolemia. In the right time, proteins are temporally dynamic. They can give us a snapshot of disease.

A much more elegant and concise way of framing this comes from my friend Alyssa Miller and Eric Fauman of Pfizer. They say that proteins speak louder than genes, and ultimately, I think we could use proteomics in a much more powerful way for precision medicine than we could for genomics.

Before we go any further, I might want to stop and just explain what proteomics is for anybody who is curious or a novice to this field. The oversimplified explanation of what proteomics is, it's obviously the study of proteins, but we could broadly break it down into targeted proteomics in which we have an immunoassay that's designed to typically detect one protein and detect that protein very well.

Now, of course, this is great if we have a particular biomarker already in mind, but if we want to do discovery work or if we want to model multiple pathways at the same time, then we might want to consider multiplex proteomics in which we measure many different proteins at the same time.

You could further subdivide multiplex proteomics into mass spec and affinity-based. Of course, mass spectrometry has been around for a long time. It's an unbiased and versatile technique. However, it can be slow, it can not be particularly high throughput, and it's not particularly sensitive to lower expressing or lower abundance proteins. That's where we move into the affinity-based approaches, the topic of today's discussion. We have antibody-based and aptamer-based. Of course, there's been a lot of work with the aptamer-based at SomaLogic approach, and you'll hear more about that over the coming slides.

Let's, now that we've covered what proteomics is, talk about how it could be a useful approach for precision medicine. Firstly, for finding the right drugs for patients. We can combine proteomic measurements with genetics to perform a technique called Mendelian randomization, which allows us to, in fair causality. It allows us to find proteins that have a causal link to disease. When we find these causal proteins, there's been evidence to suggest that they're at least twice as likely to succeed as drug targets.

This has been looked at quite systematically. There was a great paper in biological psychiatry this year, and they conducted the SomaLogic SomaScan across a few thousand people in both CSF and plasma. They looked at Alzheimer's, ALS, MS, and PD. As you can see, they were able to map different proteins as neuroscience drug targets based on their likely safety and their likely causality, which is essentially a surrogate for whether it would be efficacious or not. That's finding the right drugs. What about finding the right patients? Well, we're seeing enormous progress in this domain.

Take for example p-tau217. It's a particular form of the protein tau. We can measure this in blood, and it correlates very well with levels of brain amyloid from a PET scan. We're actually using this in our clinical trials, including a Janssen at J&J. We're using it to find people who are more amenable to the drugs that we're testing in clinic. We can also use more multiplex approaches to find subtypes of Alzheimer's and other diseases.

Betty Tijms published a great paper on this back in 2020 where she conducted multiplex proteomics run on supervised clustering on the protein data. Was able to find three distinct subtypes of Alzheimer's disease in the EMIF-AD cohort, one that was more based around hyperplasticity proteins, one that involved innate immunity proteins, and one that involved complement activation proteins. You could imagine in the context of precision medicine and drug discovery if we have a drug we're developing that might target complement pathway proteins, then we could find those particular Alzheimer's patients that have more of a complement-driven component.

Then finally, the right time, finding patients at the right time. You're going to hear from Keenan Walker in a little bit, and he's done some fantastic work showing how we can combine proteins together and build these SomaSignal Tests to predict five-year incidence of dementia so to determine whether someone's going to develop dementia five years before it occurs. There's also been some more recent work in some of the biobanks to show that we can predict 10-year incidence of Alzheimer's disease and Parkinson’s disease with pretty high AUCs of 80% or higher using a combination protein approach.

Just finishing up. A lot of the work that you might have seen has been maybe at the medium scale. Maybe a couple of thousand people measured in CSF and blood. The field is quickly moving towards the population scale. I talked about population genetics. We are now moving into the era of population proteomics. I'm sure that many of you have heard about this recently. However, some of the studies which have conducted population-scale proteomics, they've mainly relied on white individuals, so they've lacked diversity. Many of the disease endpoints are also quite shallow, so we're not going to get a mini-mental state score in some of these cohorts that have already been profiled.

That is where this new consortium comes in, the Global Neurodegeneration Proteomics Consortium, or the GNPC. This is something that's being driven by Gates Ventures with support from Johnson & Johnson and many of our academic collaborators. Our goal here is to bring together many of the preeminent scientists in the neuroscience field who have been conducting multiplex proteomics and build the world's largest data set for neurodegeneration in the proteomic space. Of course, you'll hear now from Carlos and then after him, Keenan, who are both participating in this. With that, I think I'll hand over to Carlos.

Dr. Carlos Cruchaga: Hi, this is Carlos Cruchaga from Washington University. Today I want to talk about some of the research projects we have in the lab around using proteomics in order to understand the biology of Alzheimer's disease. Alzheimer's disease is the most common native disease characterized by the presence of extracellular A-beta aggregates or plaques and intercellular deposits of tau from tangles. From a clinical point of view, Alzheimer's disease is characterized by memory loss, change in behavior, and difficulty in solving problems.

Multiple studies, including genetics, proteomics, and transcriptomics have really been instrumental to identify new biomarkers and some of the current targets that are being tested in clinical trials. However, we still have not solved the disease, and additional studies are needed in order to identify better treatments.

In my lab, we are focusing in performing deep molecular profiling of human samples in order to really identify novel and risk-protective variants, create individual-level prediction models, and identify drug targets. We are doing this by generating high-throughput, unbiased proteomics, epigenetics, metabolomics, tracheotomies, and lipidomics in tissues that are relevant to the disease, including brain, plasma, or cerebrospinal fluid. We are doing this in individuals that have the sporadic from the disease, Mendelian forms, or individuals with risk variants.

We have put a lot of effort in doing this deep molecular profiling. We have generated proteomic data in more than 1,000 brain samples, almost 3,500 CSF, and almost 6,000 plasma samples. We have also generated data, all the type of omic data, including methylation, metabolomics, and transcriptomics. Obviously, today I'm going to focus on proteomics, which is some of the most advanced analysis we have done.

With the proteomics, we have been using the SomaLogic 7K in, as I mentioned, CSF plasma and brain, and we have used this data in two different type of analysis. The first one, we have integrated proteomics with genetics in order to identify genetic variants that modify protein levels or that regulate protein levels. Those are called protein QTLs. Then we have integrated those protein QTLs with colocalization, PWAS, Mendelian randomization in order to identify causal and druggable targets. We have also been using that data in order to perform more classic biomarker discovery analysis in order to identify proteins dysregulated in Alzheimer's disease. Today I will cover these two very briefly.

The first part is when we integrate genetics with proteomics in order to identify these pQTLs. Very briefly, we've done this for proteomics but also for metabolomics in brain, CSF, and in plasma, which we have done also a multi-ethnic analysis, performing analysis in non-Hispanic whites, as well as African-Americans. We have identified thousands of novel pQTLs. Some of these are tissue-specific, as I will explain, but also some race-specific pQTLs that I will not have time to focus today, but please ask me questions. I'm going to focus initially on the CSF pQTL atlas because, as you can see, we have the largest numbers of findings.

In this study, we perform a discovery and replication in which we analyze around 1,500 samples in each state that we use for validating our findings. This is a Miami plot because it looks like the profile of Miami and reflection in the scene. Don't represent **[unintelligible 00:16:13]** genetic variant, which is a chromosome and based by position. In order to be considered significant, this needs to be higher than 11. As you can see here, we have a very significant association in this analysis, and you will need to believe me that there are two different Manhattan plots, but they look the same because this approach is high power and high replicable.

In general, as I mentioned, we identify and replicate more than 3,000 independent proteomic QTL signals. Similar studies have been performed in plasma, even larger sample size.

The next question was to determine whether the CSF proteomic QTLs that we identified were specific to CSF or has been already reported. We used some of the largest studies, as I mentioned, and we found that around 500 of our pQTLs has been already reported in CSF, and most of those are cis-pQTLs, so those that are close to the gene that qualify the protein. We identified that around 1,100 of our CSF pQTLs are tissue-specific, meaning that that protein was measured in plasma, but the signal was not identified.

We also have around 600 PQTLs that are novel because the protein that we measured in the 7K has not been included in previous analysis. In general, we have around 75% of the CSF pQTLs have never reported in any study so far.

The protein QTLs are just one type of pQTL. The most common types of QTLs are the expression QTLs. We can use RNA-seq in order to identify genetic variants that are associated with RNA levels. Other types of QTLs include splicing QTLs, DNA methylation QTLs, or histone modification QTLs. Then we want to compare or determine whether the protein QTLs that we identified are also similar to those that has reported by other QTLs types.

We did this by tissue. What we found is that a very large proportion of the protein QTLs in CSF more than 70% of the protein QTLs do not colocalize or do not overlap with any other types of QTLs, including methylation, expression, splicing, or histone modification. This is indicating that protein levels are regulated at different levels than just gene expression. That includes cleavage, binding to the receptors, phosphorylation, and many other post-translational modifications. This data clearly indicates that if we really want to understand the biology of complex traits, we need to go beyond just expression and include other types of QTLs. In this case, you can see protein QTLs are really not well covered by gene expression.

Now that we have this very large and comprehensive atlas of pQTLs in CSF, what we did was to integrate this atlas with the largest GWAS for Alzheimer's disease that was reports so far, they include almost 500,000 people. We use genetic approaches and new statistical models to identify proteins that are genetically dysregulated in Alzheimer's disease. We found that there were 473 of these proteins. We also use Mendelian randomization in order to identify proteins that are part of the causal pathways. We identify 37, and then we also use co-localization in order to determine what genetic variants are associated with both Alzheimer's disease risk and protein levels at the same time.

When we put all these three analyses together, we identified a subset of high relevant proteins that are significant in at least two of these analyses. These analyses as they are combined in multiple statistical approach are prioritizing proteins that are causal and also draggable.

If we analyze what type of proteins we are identifying in this analysis, we see that these proteins are enriched in Alzheimer's disease, which is a nice positive control, brain atrophy, but also other pathways that we are also known to be implicated in Alzheimer's disease. One of those is the regulation of the immune response. We identified some of the proteins that we already knew, like CD33, or TREM2. We also identified novel proteins that have not been associated with Alzheimer's disease before, like CR2 or SIGLEC9.

When we also take a look to which proteins we identify, now we can start putting all these proteins together as part of the same pathway. Again, because we are using genetics and mediation analysis, we know that these proteins are not just a simple association, and they are part of the causal pathway.

Another interesting pathway that we identified and reached for these proteins was the lysosomal, bundle lysosomal pathways, including non-proteins like granule or TMEM106B, but other new proteins associated with Alzheimer's for the first time, like CLN5 or CTSH. Again, we now can start putting together all these proteins in the same pathway. We want to identify new therapies. We know that we can target not only these proteins but this pathway in general.

Another example of how this data integration and this bias approach is really important to understand the biology of the disease is if we focus on a specific protein, in this case, TREM2. TREM2 is a non-AD gene risk that was identified several years ago, but we still don't know the full biology of TREM2, or all these genes is leading to disease, but other genes are part of the TREM2 pathway.

We perform GWAS for more than 3,000 individuals, and we have identified four different low size for Alzheimer's disease. One is in the MS4A region. The one is TREM2 itself. This is a C signal. Another signal in chromosome 3 and another in APOE. The MS4A is a signal that has been reported in Alzheimer's disease many years ago, but when this signal was associated with MS4A, we didn't know what was the mechanism by which MS4A modified risk for Alzheimer's disease. This study alone is already indicating that MS4A modifies risk for Alzheimer's disease by modifying TREM2 biology.

This finding alone is super important because now we have a biological context for MS4A. This finding has been only...be able to lead to this finding by using these high throughput and bias analyses in human samples.

In collaboration with Celeste Karch and Laura Piccio and others, we were able to demonstrate that we, in fact, can regulate or modify soluble TREM2 levels by modifying MS4A4A by over-expressing and knocking down MS4A in primary macrophages. The signal chromosome 3 is also a very interesting signal that has been-- This is also a new finding that includes two genes in the region. One is RBMS3, and the other is TGFBR receptor 2.

This is the first time that these gene are be linked with TREM2. Similar to the analysis in MS4A, we use primary macrophages to over-express and knock down these two candidate genes. We found that TGFBR receptor 2 and no MS4 and no RBMS3 also modified soluble TREM2 levels. We are now identified novel gene part of the TREM2 pathway. We can put all those genes in as part of the TREM2 pathway. We identified MS4A for the first time. Now we identified TGFBR receptor 2 and also APOE. All of these potential new genes could be new therapeutic targets.

[inaudible 00:25:45] This is an example of how integrating genetics and proteomics, we can provide a biological context of some of the genetics due to one's low size but also identify high-valuable targets. We can also leverage this proteomic data to identify biomarkers for Alzheimer's disease. In this case, we leverage the proteomic data that we generated in brain, in CSF and plasma, for control individuals that have Alzheimer's disease and also TREM2 variants. In this case, we wanted to identify these specifically proteomic signatures of TREM2 risk variant carriers. We have between 100 to 21 TREM2 risk variant carriers. This number is quite large because variance in TREM2 has quite a very low frequency.

In order to do this, we leverage the data to first identify proteins that are dysregulated in TREM2 compared to controls or in TREM2 compared to Alzheimer's disease cases in both CSF and plasma. Then we replicate these proteins, and we leverage those proteins to create prediction models. We identify a subset of between seven to nine proteins that have a very high predictive power to differentiate TREM2 for controls or from Alzheimer's disease cases in CSF and in plasma with AUC higher than 0.85.

This is important now because we can predict TREM2 risk variant carriers, because if we want to do that, the simplest thing will be just to sequence TREM2, is going to be faster and simpler. What we are doing here is the opposite. If we have individuals that are coming to the clinic that have TREM2 variants, we should use one prediction model specific from these individuals instead of using a more general prediction model or biomarkers, like CSF p-tau/Aβ42. These analyses are also indicating that even these individuals with TREM2 variants develop AD, the pathways that lead to disease are unique to these individuals.

In this study, as I mentioned, I present here the TREM2, but we performed a similar analysis for sporadic ADs and for autosomal dominant ADs. We identified some proteins that were common across all of them, a small one that has been reported multiple times. We also have some proteins that are common between sporadic and TREM2, like calcineurin and 14-3-3s. We also find proteins and pathways that are unique to sporadic Alzheimer's disease, TREM2 carriers, and autosomal dominant ADs.

This data suggests that if we really want to understand and create new biomarkers for Alzheimer's, we need to be able to understand the heterogeneity of the disease and also create prediction models in more homogeneous groups, those can include individuals with TREM2 variants, specific APOE genotypes, or other variants. In some of the new analyses that we are doing in the lab, we are moving in that direction. In this study, include some of the previous proteomic data that we generated, but now with this large, expanded data set that we have, we expect to validate and extend these findings.

Just to finalize this talk, I think I have showed you some just a sneak peek of some of the projects that we have. This is a good example of how integrating human genetics with proteomics and functional genomics is going to be instrumental to understand the biology and the specific events that leads to disease, and the MS4A and TREM2 finding is a very good example of that.

I also think that these high-throughput omic approaches as they are free of pre-conceived biases will lead to novel and exciting findings. Again, I think that the link between MS4A and TREM2 is a good example. We also have a recent paper that we performed similar analysis in LRRK2. LRRK2, which is a PD gene. We identified around 24 proteins that were dysregulated by LRRK2 variants. Again, the link of those proteins with LRRK2 was no report before. As I saw, we also can identify new therapeutic targets, and MS4A4A is now being used as a target in multiple clinical trials.

In summary, what we are doing in my lab is we are trying to do molecular profiling, not only doing genetics but transcriptomics, proteomics, epigenetics, metabolomics in tissues that are relevant to the disease in a cross-sectional and longitudinal manner using very large and well-characterized cohorts. We think that this data is really going to identify new genes and pathways. It's going to help to put those genes together as a part of a specific pathological event. We will lead or we will identify new prediction models and new therapeutic targets. We **[inaudible 00:31:29]** this data to identify more homogeneous group of individuals and get to more personalized treatment.

As I mentioned, we are generating a very large amount of data, and we are putting all this data in the NIH or NIA-approved repository, in this case, is NIAGADS. We have a specific collection for us, which is the Knight-ADRC collection as most of our sample are coming from the Knight-ADRC. We are not only sharing the raw data. We are trying to make easy to access and query this data, and we are putting a lot of effort in creating browsers to analyze this data.

We have one browser for the multi-tissue proteomics, the last part I present today. Also, we have released a new feed web that includes genetics and QTL, sorry. These include brain, CSF and plasma, proteomics and metabolomics in non-Hispanic whites and African Americans, and it has more than 26,000 molecular traits. This is a very great resource to query and analyze and explore this data without the need to apply to any data repository or download anything on process anything.

I want to thank all the people in the lab that has been instrumental to perform this analysis. Judy Wang, the new TREM2 analysis in collaboration with Yun Ju Sung. Dan has done all the CSF pQTL. Priyanka has been involved in leverage and QC the genetic data. Jigyasha, QC all the proteomic data. Many other people have been involved in selecting the samples, pulling the samples, and so on. We have a lot of data that we are now analyzing, and we have multiple positions to work with this and other data. If someone is interested, please let me know.

I want to finalize thanking all the funding agencies and, the ADRC, DIAN, Fundació ACE for sending the samples with us. With that, I will be more than happy to take questions. Thanks very much.

Dr. Keenan Walker: Okay. Hi, everyone. Thank you, Carlos, that was wonderful, extremely important resource, which I do plan to take advantage of in the near future. I want to thank SomaLogic and Nature for putting together this wonderful session. Previous two speakers have been extremely informative. I'm going to talk about prediction using proteins to predict dimension risk, future dimension risk specifically, and using this large-scale proteomic platform.

A lot of the work that I do is based on this idea that health conditions that lie outside the central nervous system, so systemic health conditions, things like diabetes, hypertension, autoimmune, inflammatory conditions can drive dementia risks or at least influence the risk. This figure on the right is published by the Lancet Commission, and I included it to illustrate that anywhere from 30% to 40% of attributable risk for dementia is linked to conditions that lie outside the central nervous system that are also modifiable. We believe that proteins in circulation might actually mediate this relationship between systemic health and Alzheimer's disease and dementia risk.

The framework is provided up top where we think that whether it be disease, subclinical conditions, or non-disease factors, like cellular aging, drive abnormal protein expression across various tissues. These proteins that make them weigh into the bloodstream and through various conduits can influence cells within the brain. A number of these proteins have been identified previously, a number of cytokines especially listed here, but we believe there's many, many more systemic proteins involved in Alzheimer's disease and dementia risk that might be able to inform dementia prediction, but also may be viable therapeutic targets as we just heard.

Our team has done this for some years now really with the advent of large-scale proteomics. We've particularly gotten a lot of use as a SomaScan platform. In 2021, we published a paper *Nature Aging* showing that a number-- we identify a large number of proteins that were novel that also predicted dementia risk over a five-year of follow-up period. A subset of these proteins were identified to be potentially causal, too, which was particularly interesting, but this was in a group of older adults. We recognize that although a lot of the proteomic work has been done in older adults, Alzheimer's disease and other forms of dementia really do start multiple decades before late life.

We know the preclinical period of Alzheimer's disease is very protracted, it takes place over two to three decades before the expression of symptoms where you see pathology on the brain that's progressing. There's this consensus now that

Alzheimer's disease and other dementias really begin as early as midlife. We sought to look at midlife to try to understand what proteins might be influencing dementia risk or related to dementia risk to identify potential mechanistically relevant proteins and also potential biomarkers as well.

This study was published recently in *Science Translational Medicine*, where we use data from the atherosclerosis risk and community study or the ARIC study to look at 5,000 proteins and 11,000 individuals who are non-demented in midlife, and so we looked at proteins one by one in relation to 25-year dementia risk. In that time span, we had almost 2,000 incident dementia cases.

Here's the first volcano plot from our analysis, and these are analyses that are adjusted for demographic confounders, cardiovascular risk factors, et cetera. What we found was that 26 proteins were associated with 25-year dementia whereas when they were measured at midlife, and these covariates, potential confounders were adjusted out. GDF15 you'll see rose well above the others up here.

We also looked at the same set of 5,000 proteins, this time in relation to what we call near-term dementia risk or dementia occurring within 15 years. We found GDF15 again popped up, but also six new proteins that weren't shown in the primary analysis. We have this additional set of proteins that seems to be more so an indicator of incipient dementia, even when measured during middle adulthood, and we did the same thing for dementia beyond 15 years. Although these proteins were identified in the full follow-up time, we can say that these proteins especially are altered well, well before the onset of symptoms. 15 years, so likely as much as two decades before the onset of symptoms.

When we put everything together, we have 32 midlife dementia-associated proteins that we identified across the various time spans. GDF15 was the one that stayed consistent throughout all the follow-up windows. We have a subset down here on the bottom left that is short-term-- 15 years, not so short, but more near-term specific. Everything that has a star has been nominated previously through the accelerated medicine partnership as a potential therapeutic target. We do see some convergence here looking at the midlife proteome in blood, which was really exciting to see.

We asked the question, are these actually bona fide midlife risk factors? Our sample mean age was 60, but we said, oh, we're big enough to cut the sample in half. Let's look at people in the 40s and 50s, run the same 32 proteins over the same follow-up periods, and see how associations change.

We see original in blue, our younger half in red, and essentially things stay the same. We really do think we have fourth and fifth decade of dementia-associated proteins here, truly midlife. When we look at the biology of these 32 proteins, they fall into some of the categories that you've already heard about today. Immune function, proteostasis, synaptic function, and several others. Of course, we wanted to validate our associations and we were able to do so in collaboration with investigators from the European Medical Information Framework and the Whitehall-II study. We saw a subset, not all, but a subset of our proteins were validated in these different study designs. It's not exact validation, but it does give us quite a bit more confidence in the number of the candidate proteins we identified. What was really interesting and what we were allowed to do in collaboration with the EMIF cohort was look at how our individual proteins related to neurobiologic changes as measured by CSF biomarkers of amyloid beta, phospho-tau, total tau, to really understand, are we looking at Alzheimer's relevant processes or something more non-specific.

You can see a subset of our proteins do show associations with ADRD, well, Alzheimer's disease biomarkers specifically. Then we look at total tau, we look at NfL, neurogranin for markers of neurodegeneration and see some additional proteins, which are linked to CSF markers of neurodegeneration. Then we also look at YKL-40 to capture neuroinflammation. Same thing. I do want to turn your attention to GDF15, which we do not see, which was our strongest in dementia-associated protein linked with dementia risk across all follow-up windows.

Both replication samples confirmed this, but it's not associated with amyloid, not associated with p-Tau, but it is associated with our neuroinflammatory marker YKL-40. We can characterize how these proteins, if indeed mechanistic or signals biomarkers of what's going on in the CNS, what they tell us about the neurobiologic changes, with which they're associated. We were also interested in where these proteins are coming from, and we were able to use open-source data to do that. GTEx, Human Protein Atlas. With that, we can understand that while most of our proteins, a lot of them were non-specific, they came from many, many tissues, not just the brain.

We did have another group, however, a smaller subset of our proteins that are CNS-specific. Those are synaptic proteins, as you might imagine, complexin-1, complexin-2, cerebellum 4. We had a number of these which do seem to be seen as specific. When we do pick them up in the periphery, we were confident that they're coming from the brain or other processes within the CNS and may make especially good biomarkers. We then have this other set of proteins that doesn't seem to be expressed in the CNS at all or at measurable levels.

That includes things like GDF15, which based on these databases, doesn't seem to have very high levels of expression in the CNS, despite its strong, strong association with clinically relevant outcomes. For those proteins that were in the CNS, we can use open-source data to understand if they are differentially expressed in Alzheimer's disease brains. Just to summarize this, the stress response and immune proteins seem to be upregulated, and the complexins, the synaptic proteins seem to be downregulated in the AD brain. Actually, this association reversed for the synaptic proteins. Going from midlife plasma to late-life autopsy brain tissue, we see a reversal in the synaptic signal, but still a strong associations.

Then we asked the causal question. We used two sample Mendelian randomization to do this in collaboration with Myriam Fornage and Yunju Yang, pictured here. We identified the pQTLs, protein quantitative trait loci in our cohort, and we used them as instruments to see if the proteins might be causally linked to Alzheimer's disease. Of the 32 we looked at, we found evidence for calsyntenin-3 and SERPINA3, showing these nominal associations that suggest a causal link. SERPINA3 in particular was robust to sensitivity analysis. We're more certain about that relevance. That's to be determined. We really want to pursue mechanistic understanding of how SERPINA3 might be apparently protective.

What was more striking was in the reverse direction. When we look to see if Alzheimer's disease or Alzheimer's disease genetics drive protein expression, we see about half of our proteins are differentially expressed, seemingly as a result of Alzheimer's disease. Based on our study, we can say that they're differentially expressed as early as the fourth of the decade of life. In addition, this can be seen as reverse causality for these proteins, and I think that might be some of what we're seeing. We also have some potentially genetically validated biomarkers of AD.

I'm not saying they're specific to AD, but it seems to me that Alzheimer's disease might be driving differential protein expression in plasma as early as the fourth decade. All in all, we have this new set of midlife biomarkers that represent a more diverse pathophysiology. We're going beyond amyloid, going beyond tau, and we think we have some indicators of proteostasis breakdown, immune dysfunction or protective immune function, or lack thereof, synaptic function and other pathways like vascular and ECM processes that are known to be relevant to dementia risk.

What about prediction? These proteins, when trying to predict 25-year dementia risk from a midlife period, they don't perform so hot. We're seeing AUCs proteins alone about 0.66, demographic factors, which is quite a bit, age, all the things listed down here, education, plus clinical risk factors gets us into the mid 0.7s. Then when we add proteins on top of that, we get a little improvement in our predictive accuracy. We're getting close to 0.8. That improvement is significant, though somewhat marginal. When we look at prediction of long-term dementia risk. I should say we're using only the top candidate proteins to do so.

We started with the 32, now we're looking at the near-term proteins **[unintelligible 00:46:56]** seven, we see proteins alone gets us a AUC of 0.67. Demographic factors plus clinical variables gets us near 0.8. Then you get a slight improvement by adding proteins. Still, proteins alone, not so predictive of long-term dementia risk, at least in this context, and lots of reasons, our outcome is very heterogeneous.

While most of it is likely Alzheimer's disease or mixed AD, vascular, neuropathology, there's lots of other stuff in the mix that will inevitably put a cap on how well we can predict outcomes. Again, we're looking over a 25-year time horizon, but we really focused in to see, can we actually improve our prediction if we really transition from focusing on understanding biology to trying to optimize midlife dementia risk score to predict future dementia risk.

That's the second half of what I'm going to talk about today, where we identified a combination of biomarkers that we think to be most predictive, or that we found the most predictive of dementia when measured during midlife. This is now referred to as the Dementia SomaSignal Test. Actually, this test is available for research use through SomaLogic. This is an effort that was led by Clare Paterson and Michael Duggan. Clare works with SomaLogic, she's R&D. Michael works -- he's a postdoc in my laboratory, and they've both done some really fantastic work with the rest of the team to develop and test this dementia SomaSignal Test or midlife dementia risk score.

We used the ARIC study for this. We divided the cohort. Importantly, this is a community-based cohort, not clinic sample. This is representative of at least Black and white individuals within the United States. We divided the sample in 70, 15, 15 for training, tuning, and validation respectively. Then we did secondary validation where we looked at the risk score in late life and related the risk score to other EDRD phenotypes. I'm going to walk you through that, but to develop our score to select proteins, we first looked at the univariate association.

Not considering our confounders here because we're just interested in protein prediction of an outcome, but we took the top 50 proteins from our univariate association and then ran those through machine learning, particularly elastic net with 10-fold cross-validation. From that, we got this combination of 25 proteins that was ultimately predictive of 20-year dementia risk. We have this dementia risk score and we computed the absolute risk. Then we bend people into four categories, low, medium low, medium high, and high risk. We'll talk more about that here.

In terms of just predicting or using it continuously to predict dementia risk, we saw AUCs in our training model in the low 0.7s and then when we did our validation set, which was the 15% holdout, we saw AUCs slightly lower, but still at 0.7, actually exactly here. We compared that to how well E-four status predicted a 20-year dementia risk just as a comparator, and so we do outperform that. We looked at other comparators as well. I'm going to get to that in a second, but first I'm going to talk about the four bins.

Because we want this to be in the future clinically relevant, we thought, "Okay, let's bin people into these four groups and see what the risk over this 20-year period of dementia is." We see the high-risk group is at much greater risk for dementia over this 20-year period than the low-risk group and we see the stepwise increase in the intermediate groups. When we look at the risk ratio, the high-risk group is over nine times more likely to develop dementia over a 20-year follow-up period than the low-risk group. We have a pretty strong stratification there.

We also were interested in comparing midlife dementia risk score to biomarkers that are commonly used at least in research right now. We used what we're calling the Canonical AD and Neurodegeneration Biomarkers. We're all very familiar with them, like Aβ42, GFAP, NfL, or pTau-181. This is from the Quanterix platform. We were able to, in a subsample, compare how well our dementia risk score, which we're calling dSST here, predicted dementia over 20 years compared to Aβ, pTau, NfL, and GFAP, and so it did outperform all of those coming in at about 0.7. It also outperformed the combination of these four biomarkers.

I'll admit, these four biomarkers aren't designed to necessarily predict dementia. They're designed, well, at least Aβ and pTau, to identify specific pathology, so to predict an etiology. In some ways, it's an unfair comparison, but there's still an unmet need for actually prediction of clinical outcomes. We see that our risk score does do that at least better than what's out there right now.

We also see that when we combine the Quanterix biomarkers with dementia SomaSignal Test or dementia risk score, we perform even better. We show in a separate sample of the Baltimore Longitudinal Study of Aging that having higher levels of dementia risk score associated with a future cognitive decline being in this high-risk bin is associated with cognitive decline across every domain that we examined.

We also show that this risk score is associated with lower brain volume, but also longitudinal brain volume loss across lobes and within gray and white matter pretty consistently. When we do DBM to try to localize where the strongest effects are, we see medial temporal associations. Although we're predicting all-cause dementia, to some degree, we're seeing prediction of atrophic patterns that tend to be observed in sporadic AD. We also showed that our midlife dementia risk score is associated with, again, canonical Alzheimer's disease and neurodegeneration biomarkers in the way we might expect, lower Aβ42 to 43, higher GFAP, higher NfL, and higher pTau-181 and we see this sort of stepwise increase across the bins.

Lastly, we were able to show, well, we examined at least how our midlife dementia risk score is associated with Cortical Amyloid. Now we're talking about etiology. We see that yes, there is some separation here. The dementia SomaSignal Test is predictive of elevated amyloids. We see a strong odds ratio here, but when we compare it to things that are designed to predict amyloid status, likely Aβ, pTau-181, it underperforms slightly in these measures, well, more so a larger gap between it and pTau-181.

We do see, and this is to be expected, but when we add our Quanterix biomarkers, Aβ, pTau, NfL, GFAP to our dementia SomaSignal Test, we get AUCs approaching 0.9 and with improvement that's not negligible by including the dementia SomaSignal test. These can be used in combination to try to predict amyloid status in people without, for example, amyloid PET scans or CSF data available.

Just to summarize, we've developed this 25-protein composite score that's predictive of all cause dementia in a 20-year follow-up in a community-based sample. That last part is important because the prevalence of dementia is lower than what we might think, so achieving a higher AUC is harder in this context. We see superior classification to widely used markers of Alzheimer's disease pathology, neuronal injury, and astrogliosis, and the dementia SomaSignal Test or midlife dementia risk score is predictive of cognition, cognitive decline, cortical atrophy, and cortical amyloid pathology.

Lastly, there are some implications and potential uses. It can serve multiple purposes, including to help enrich trials, identify people who are actually going to experience clinical progression once we understand what's going on from an etiology perspective. It can also help reduce screen failures and again, monitor change over time, potentially as a proxy outcome. Again, thank you for the opportunity to speak, and thank you for listening.

Sarah: Okay. Thank you very much to Keenan, Carlos, and to Chris for those very informative presentations. We will now move to the question-and-answer session. We'll stay with you an extra five minutes, so we have time to answer all the questions, so do please remain with us. To ask a question, please type it in where it says type your questions here, and then press submit.

All right, let's move back into the studio and I will begin to ask questions. I'm going to begin with this question which came about halfway through Carlos's presentation, actually, on data and what a rich resource of data we are now developing. Will these genome-wide studies, proteomics, lipidomics, metabolomics data be made publicly available? If so, what's the timeframe for that? I'm going to probably direct this to Carlos first since that is where the question arose.

Carlos: I think these are great resources and they should make available as soon as possible. We are making an effort to make available both the raw data as well as the process data and the process. From our point of view **[unintelligible 00:57:29]** data is already uploaded to the NIH-approved repository called NIAGADS**,** so all the data is there already. Then we are making an extra effort for the process because I think that is helping the community a lot.

Sarah: Absolutely. Thank you so much for clarifying that, Carlos. Our next question asks about the interplay between the blood and the brain. What proteomic interplay is expected between the central nervous system and the periphery? I want to direct this to Chris. I can't actually see you, Chris, at the moment. I don't know if you're using two screens, but currently, I can't see, although I can hear you.

Christopher: Yes, I'm not quite sure what's happening with my camera. I apologize.

Sarah: We can hear you though.

Christopher: Yes. It's on. That's a great question. We've explored that exact question using data from both of the major affinity-based platforms, and the results are pretty similar. For example, we ran the SomaScan 7K on both spinal fluid and blood plasma samples from a cohort of about a thousand people in Barcelona. Of the approximately 7,000 **[unintelligible 00:58:46]** that were measured in both CSF and plasma, the correlation is high with the sort of an r-squared of 0.7 for the first 60 proteins. That's less than 1% of the proteome showing a high correlation. Then there are about 420 showing moderate correlations of 0.4 or higher, so 6%. For the remaining 93% of proteins, the correlations are low.

There's a couple of caveats here that are important to point out. Firstly, it's hugely important to highlight, the findings I just mentioned, they're not peer-reviewed, they're in need of further testing. Second of all, though, it's worth noting that those results are based on total protein concentration. If we take pTau-217, which I mentioned earlier, if you look at total tau levels, they don't necessarily correlate well between blood and CSF, but pTau-217 on the other hand, that's a very promising blood-based surrogate of brain tau and amyloids. It's important to keep in mind the proteoforms and the PTMs when we talk about the interplay between the periphery and CNS. Apologies for the camera. I don't know what's going on.

Sarah: Okay, thank you. I don't know if either of Carlos or Keenan, you have anything to add to that?

Keenan: Yes, I can add to that. That's going to be a great resource, Chris. Wonderful. I just want to add that in order for protein in the periphery to have some effect on the CNS, it doesn't necessarily need to make it into the CNS, or it doesn't necessarily need to be a strong correspondence to the same protein in the CNS. We know of proteins that it's a lot of the cytokines, a lot of the chemokines that through signaling things like the endothelium, can actually influence neuroinflammation, for example, or the aggregation of-- or these **[unintelligible 01:00:42]** aggregates without ever even getting into the CNS.

They do set off a chain of events that-- Ah, there he is. They do set off a chain of events that does affect the brain in a meaningful way. I think knowing the CSF plasma correspondence is going to be tremendously valuable.

Christopher: Absolutely.

Carlos: Yes. I also want to reiterate the same. I think that we have also data that is still not **[unintelligible 01:01:14]** in brain and CSF on plasma, and we see higher correlation between brain and CSF, which is expected, and lower with plasma. In any case, I think that the results from Keenan is very interesting. I think there are going to be biomarkers in plasma that are all going to have a subset of different proteins to that of CSF. Some of this are going to be overlapping, but clearly for biomarker and prediction models, their field is moving to plasma, and I think we are going to get good results.

Sarah: Okay. Thank you all very much. I'm aware of time, and so I'm going to jump to the next question, which is perhaps best made now so that we include it. I just want to ask how you plan to bring these discoveries to clinical applications, either in trials or in healthcare settings? Could you speak to that? Anyone can open with that. [laughs]

Keenan: Because I think there's less daylight between bringing biomarkers to clinical practice than there is bringing disease-modifying therapies, I'll go first just with a low-hanging fruit. The midlife dimension risk score, although it's marketed for research use only, we know people are interested in it, specifically in certain regions of the globe, in understanding their dementia risk at midlife, for better or for worse.

I'm not sure if I necessarily want to know, but we think people can use it in the near future to understand what their risk may be, and maybe enhance their motivation to start making changes, especially things like diet, exercise in a way that can perhaps lower the risk. We'd be able to use the midlife dementia risk score to monitor risk over time and the influence of interventions on dementia risk. I don't think that's far off. Of course, there'll be combined with etiological biomarkers like amyloid and pTau, but I think in combination you can get a lot of insight in the very near future.

DeAunne: Yes, I'd just like to add to that very quickly. SomaLogic actually has a test in clinical use for this in Japan currently. That's how close we are to direct clinical application.

Sarah: Thanks very much for that, DeAunne. Carlos or Chris, would you like to speak to that, or should I move to our last question?

Christopher: Let's see the last question.

Sarah: [laughs] I'll move to our last question, which is a bit of an all-encompassing one. Any of you could take it in any direction. What are the most important outcomes you have achieved thus far? I'm going to direct this first to Chris.

Christopher: That's a fantastic question. Because I work in an industry setting, some of the best outcomes I can't talk about, because we're working on it. I would say on a much broader level, we're absolutely finding new drug targets. As I mentioned, I think that those are going to be a lot more likely to succeed. We won't bear the fruits of that yet. It takes so long to develop drugs. It probably won't be for another few years until we see those outcomes. I think that the predictive tools are one of the most exciting things that are going to come out of that. Keenan's already talked about it. I think that over the next two or three years that's going to develop pretty quickly.

Sarah: Fantastic. Carlos, would you like to have the last words here since I've heard from Keenan about the potential clinical applications that he feels are important?

Carlos: Yes, I think we are in a very exciting time with all the clinical trials and all these proteomic data coming online. Clearly, a lot of the clinical trials are focusing on Aβ and tau, which are also very good biomarkers, but if we are targeting those proteins, we need to develop biomarkers that are independently of those proteins.

I'm very excited to see what's going to happen in the next years. The data that is being generated by many groups is amazing, high quality, and I think we are going to be able **[unintelligible 01:05:55]** new biomarkers and new targets with these approaches. It's very good to research these days.

Sarah: All right. Well, on that note, I think we are going to have to finish here for today. It's been great to have you all. It's fantastic to hear about all the work that you're engaged with. Very exciting, I think, for us to hear, and to hear you share actually across industry, academia, clinical application, et cetera. It's really great. We are aware that there are some questions that we didn't manage to get to in the audience, and we will get back to you offline after this broadcast finishes using the details that you've provided to us, if you've agreed to be contacted.

I would like to thank Dr. Christopher Whelan, Dr. Carlos Cruchaga, and Dr. Keenan Walker very much for being with us today, and for answering all of our questions. Thank you to the three of you. I would also like to thank our webcast sponsor, SomaLogic, and of course, you the audience for taking the time to be with us today. Remember, you can watch this webcast again at any time on demand at nature.com/natportwebcasts. Thanks for watching, and I hope you'll join us again soon.

[01:07:19] [END OF AUDIO]