Rebecca: Hello, everyone. And thank you for attending today's webinar, *Harnessing AI and Proteomics for Glioblastoma*, presented by SomaLogic. I'm Rebecca, and I'll be moderating this webinar. I'd like to start off by introducing today's speakers, Andra V. Krauze, physician, early investigator, radiation oncologist at NCI and NIH; DeAunne Denmark, Senior Director of Translational Medicine and Clinical Development at SomaLogic. You can read their full bios on the left side of your window by selecting the speakers tab. Just a few notes before we begin. You can access closed captions from the bottom right corner of the video player. This webinar is being recorded, and it will be available to watch on demand within 24 hours. We'd love to hear from you during the presentation. Please submit any questions you have using the questions and answers tab on the left side of your screen. Time permitting, we will conclude with a Q&A session.

Let's begin. Andra, please go ahead.

Dr. Andra V. Krauze: Hello, and thank you for having me for this webinar. Today I'll be speaking about harnessing artificial intelligence and augmenting computational learning to improve glioma outcomes. This is the proteome edition. I will be focusing heavily not just on harnessing artificial intelligence and computational learning, but how do we go from having proteomic data to actually linking that to clinical glioma outcomes, and how we can take advantage of computational learning to do this.

With that, I'll begin. I have no disclosures, but I will acknowledge that I have a bias, and that is that I believe in single-minded dedication to harnessing growing data in the oncology and medical space to improve patient outcomes.

With that, the work we do. I am a radiation oncologist, and I work in radiation oncology branch at National Cancer Institute, NIH, and I'm also a physician early investigator. So I carry out both clinical work and seeing patients and treating patients on clinical trials, as well as carrying out academic research. We treat patients with radiation therapy. I'm a radiation oncologist. That is my job. And we run clinical trials. The goal of my research is to develop artificial intelligence methods for the curation of data, to both create an infrastructure for robust data collection in oncology, as well as harmonization of the data that renders it analyzable.

The ultimate goal is, of course, to identify biomarkers that can ultimately be linked to radiation therapy in a manner that will allow for clinically meaningful artificial intelligence-based algorithms to connect radiation oncology dosimetry data to omic datasets. Today I'll be discussing disease site brain and histology glioblastoma.

The motivation for our work, for my research, is grounded in the fact that outcomes in glioma remain poor. In glioblastoma, this is essentially a terminal diagnosis. This is multifactorial, but it is in great part due to the fact that glioblastoma functions via biological pathways that are both redundant, as well as highly heterogeneous. We have, as a result, a limited understanding of the biology and triggers that drive resistance in this disease. We do believe that large-scale proteomic data, omic data in general, and I started by saying data in the oncology space can help us examine biological response.

Given the large amount of data in this space, we do require artificial intelligence that is clinical driven and clinical curated to help identify patterns that help us then examine, understand, and ultimately target biology.

I broke up this talk into several sections in order to organize how we think about this problem. I'll start by speaking regarding glioma and radiation oncology, and then move on to the proteome, and eventually to future directions on how we can take advantage of large-scale proteomic data in this space.

I'll start by giving us some context in the clinic. This is a hypothetical fictional case of a 33-year-old gentleman, previously well. He's a practicing dentist who's presenting with a first-time seizure. He is seen by our neurosurgery colleagues. He's getting taken to the operating room, where he undergoes an incomplete resection. Given the large burden of disease you can see in this patient's MRI, this is not surprising. As well as the involvement of eloquent areas, pathology comes back as a glioblastoma, and this patient now presents to a neuro-oncology clinic. Now, the plan is for six weeks of concurrent or at the same time radiation given with chemotherapy. The chemotherapy being oral agent temozolomide. This is then followed by 6 to 12 months of adjuvant chemotherapy.

Given all the standard of care management, the actual median survival is 14 months. In more recent studies, we have been able to see higher parameters for this up to about 20 months. But generally, this is a terminal diagnosis with significant **[unintelligible 00:04:59]** and impact on the patient, both from a neurological standpoint and obviously from a survival standpoint.

A hypothetical conversation here is important to ground what we are going to discuss today. This is the filter through which I would like you to look at this talk. These questions are questions that I get most often in the clinic, and I structured them as such.

The patient usually will ask, "How will I be treated?" I already mentioned that the standard of care management in glioblastoma is concurrent chemo and radiation. Radiation together with chemotherapy given at the same time. This is standard of care, and it has been shown to improve survival. In fact, the concomitant or concurrent administration of radiation with chemotherapy is the only proportion of our treatment that has been shown to improve survival in this disease by getting radiation together with chemo.

“How will you target the radiation or the chemotherapy based on my tumor biology?” At this point, we have to acknowledge that we don't currently actually alter either intervention based on the tumor biology in glioblastoma.

Very often then, patients go on to ask, "Okay, but what about targeted agents for cancer, and which agent are you using?" We don't currently have approved targeted agents for this cancer. "How do you know that this treatment is working?" Well, we will do an MRI of the brain one to three months after completion of treatment.

"Will this MRI show that I am better or that my tumor is responding?" This is a complicated question and answer, because the MRI shows both the impact of treatment or treatment effect, as well as potential tumor response. It's difficult to be able to separate these two features until much later as we're looking sort of test of time of how the MRI evolves or potentially where we say the needle of truth, meaning we had to do-- We went into neurosurgery and obtained tissue that shows existing active tumor.

"Is there a blood test to see if my tumor is growing?" We do not currently have a test that can provide this answer in glioblastoma. "Is my treatment personalized?" It is in fact actually not personalized in the sense that although the radiation oncology proportion of the treatment is based on targets that we generate given the patient's imaging, so that component is in fact personalized, obviously, to the patient's anatomy and their tumor. It is not personalized to tumor biology, or to the individual response to radiation therapy or chemotherapy, for that matter.

Although the treatment is standard of care, this is a treatment for all patients. This combination of radiation and chemotherapy that I described with an adjuvant or helper chemotherapy given 6 to 12 months following completion of this is based on existing evidence.

This brings us to this conundrum. On the left-hand side, we see this high heterogeneity and variability, and on the right-hand side, we see what appear to look like eggs. It's actually a dessert. It's a passion fruit panna cotta. But really, the point here is that even though we are aware of this high biological redundancy and heterogeneity, we actually treat all of our glioblastoma patients with concurrent chemoradiation followed by adjuvant chemotherapy.

Therefore, this brings me to the issues at hand. What you heard the patient, this fictional interaction, what you heard the patient ask is where we should be getting our cues from. These are very important questions.

What we learned there is that there is no biomarker for tumor response or resistance. There is no personalized treatment for this gentleman. And, ultimately, the management of recurrence is heterogeneous, and effectiveness is limited. Although he didn't really ask me about that at this point yet, but that usually happens during the course of treatment or upon follow-up.

Which brings me to a brief overview of gliomas. These are malignant tumors that are derived from glial cells or their precursors. This is the most common primary adult brain tumor. The median aged diagnosis is 64 years of age, and is more common in males as compared to females. Survival, as I mentioned, is quite poor, 6.8% at five years. Typically, this tumor is actually classified based on how it looks under the microscope, as well as its molecular features.

When we talk about staging versus grading, of course, there's no staging glioblastoma, that tumor does not spread anywhere. It takes a patient's life by infiltrating the brain. Very rarely is it documented to go outside of the brain. There are case reports of this, of course. However, the patient loses their life as a result of infiltration of the brain by the tumor, and this eventually causes this to be a terminal disease.

There is molecular classification in glioblastoma, which I am alluding to there. This is comprised of MGMT methylation status, which I will discuss a little bit more in future slides, where methylation of this promoter renders this better prognosis. And IDH mutation, which also, if mutated, is correlated with better prognosis.

Now, in the big picture of management to review, we know that what we first need to do is remove the tumor through maximal safe resection, and this gives us some tissue. Then we pursue radiation therapy and chemotherapy, temozolomide, which is taken by mouth, and these are two interventions that are given at the same time or concurrently.

Following this, the patient is followed with MRI of the brain. We usually have an MRI prior to surgical intervention, then an MRI after surgical intervention, and sometimes another prior to initiation of chemoradiation. And then we have an MRI one month to three months after they complete chemoradiation, and then every two to three months thereafter.

During the treatment, we do acquire cone-beam CTs on the linear accelerator while we administer radiation therapy, and these can capture some imaging changes. Remember, these are cone-beam CTs, not MRIs. The use of these images is as yet investigational.

I went over the big-picture management of glioblastoma here, because I felt it was important to showcase where we get data in this space. We know we get some tumor tissue, as I described, and then we get some laboratory values while the patient is undergoing chemoradiation, of course every single time after that when they receive chemotherapy. And we may or may not have, in some patients, tumor tissue upon recurrence. Many patients do not, in fact, have a second resection. Only some do. The bulk of our data, I will say, is in the imaging space, given that we have these MRIs that we have prior to intervention, and then as a result of follow-up.

This brings us to our hypothesis. The hypothesis that we can use clinical imaging endo symmetry data from radiation therapy that, when aggregated and interpreted in conjunction with omic datasets, specifically in this case the proteome, we can arrive at serum biomarkers that can result in biologic interpretation and give us a peek, if you will, or a look under the hood, to be able to figure out what is going on here, why is this tumor responding or not responding, what causes tumor failure or treatment failure.

Eventually, this can lead to adaptable prognostic, and eventually predictive and scalable, AI algorithms that can help us improve outcomes. We want to go from clinical data on the left-hand side to an improvement in outcomes. On the right-hand side, you see some Kaplan-Meier survival curves that display subsets of patients, depending on how they do with treatment. We would like to, of course, elevate those patients that currently don't do so well to doing better. The question is, how can we take data to be able to elevate these curves?

Now, I come to part two, proteomic analysis and oncology. This webinar prompted me to look a little bit deeper as to how the proteome is actually being employed in evidence-based medicine. You can see that we have over 70,000 publications for Web of Science search earlier this month that involve the proteome. Of these, 3%, or just over 2,000, are existing in the space of oncology. Of these, 312 publications involve the proteome as approached by a serum as the tissue of analysis. As you can see from the numbers there, we have increasing number of publications that involve serum and the proteome. This is yet a smaller proportion of the data overall.

While the proteome in oncology is steadily rising, and that's an important piece to note there, we know that the proteome in oncology, in terms of where the data is actually present, is mostly present in these small studies. Some of them are based on animal models or tissue culture. We have relatively little data that originates from as a biospecimen from patients on trials that are undergoing both trial treatment, as well as standard of care management. A lot of the data really lives in mass spectrometry.

When we look at datasets that have associated data in the public domain, if we look at proteomic biomarkers in oncology, you can see that datasets that have associated data that is available for analysis or validation or transferable algorithms, which I'll come to in a second, there are only roughly between 17 and 25 datasets that would fulfill this brief. A lot of these are not likely to be quite as comparable to the data I will show you today. I'm planting this with you today, because it is extremely important for us to be able to share our datasets. Of course, in a safe, respectful manner to our patients while they're identifying data to be able to put this in the public domain for people to be able to advance patient outcomes by using data, including but not limited to, the proteome.

This brings me to challenges and templates for success. Conversations often arrive at the point of, "Well, do we have any proteomic biomarkers that have actually arrived at the clinic that have actually become so successful that we are able to actually harness them in the clinic?" The reality is that there's very few, and you can see them there. These biomarkers are typically used for monitoring of the disease. They're not used for diagnosis. To a large extent, they're not really used to change what you actually do in terms of management.

The data actually lives, as I explained in my previous slide, in cell lines, animal, and tumor tissue-based analysis. This doesn't necessarily capture the tumor heterogeneity that I described earlier. Very often, it is based on relatively few proteins that are compiled as a panel versus the "proteome." Very often, they involve a single time point of biospecimen collection. These are serious challenges, because when we talk about validation and transferability, this is going to be a problem.

Further to this, the datasets are relatively small. Looking at, for example, sex differences analysis in glioblastoma, we can see that all prospective studies we can look at in the last 10 years or so really deal with a few hundred patients. Retrospective studies are much larger, but these are far less likely to have the depth of information that we would require for validation and transferability of findings. This is compounded by defining the signal with a proteome. I know that others have discussed this in webinars. This is an important point that cannot be discussed enough.

When we capture the proteome, what are we capturing? On the right-hand side, so we have the organism and we have the tumor. What parts are actually being measured? Of course, we know that when we're measuring the proteome, there's this interplay between the genome, the transcriptome, and of course the proteome, which is comprised of both altered and unaltered proteins. The altered proteins undergo several modifications, all of which are extremely important and as yet important for classification and defining the role of what ultimately happens in terms of outcomes and linkage of data.

With that, this brings me to this really huge conundrum as to when we are measuring the proteome, and we find important signals, how are we to link these bench to bedside, to animal work, to tissue culture, to be able to not only probe deeper in terms of the signaling pathways and the biology that is at play, and to be able to translate this into druggable targets?

This brings me to the part three of my talk: data types, integration, and analysis. Essentially, my work involves this very busy slide where on the left-hand side, you see the imaging that I hinted at earlier that we are capturing in patients and has been captured over years. Although siloed, it's important to be linked with radiation oncology dosimetry, as well as omic datasets. Of course, today we're discussing the proteome, which is a dataset that is as yet quite novel, but very important to be able to interrogate the biology, which we'll come to shortly.

We do know that molecular biomarkers, as you can see in the lower left-hand corner panel, are growing. It is important to be able to link this to biomarkers, not just in oncology at large, which I pointed out to you is actually still a relatively smaller proportion, because it is a very difficult area to study for the reasons I mentioned, but to be able to take this into the neuro-oncology realm. You can see that in radiation oncology, people have been analyzing biomarkers for quite some time. When we're looking at neuro-oncology specifically, this is a challenge.

In part, this is a challenge, because when we're talking tissue, of course, tissue, obtaining tissue in a brain tumor patient requires a neurosurgical intervention. It isn't simply a matter of obtaining a biopsy or getting a small proportion of the tumor, but really being able to take the patient back to the operating room, given the fact that the patient may have neurological problems and may have performance status difficulties and may really not be well enough or willing to go to the operating room to have more tissue, which brings me back to the leitmotif of this talk of serum and obtaining the proteome.

This is a good summary slide to discuss clinical deficits in oncology, where we really have this heterogeneous and imperfect capture of outcomes. By outcomes, I mean overall survival, progression-free survival, and, of course, the impact on normal tissue. With that, the ability, or often the lack of ability, to fully integrate radiation oncology dosimetry data, given that all of these data types exist in silos that typically don't talk to each other and more about this in my next slide.

When we do analysis that involves computational analysis, including but not limited to the proteome, very often when I have people start to work with me, they expect to arrive boots on the ground and have this beautiful dataset that is completely and fully labeled, and that's simply not the reality. The reality is that we spend 80% to 90% of our time going from unstructured data to semi-structured data to structured data in the repositories that you can see here, and from the electronic health record system to be able to link clinical data and the outcomes of patients to data that we measure on a large scale as may be the case in the context of the proteome, but not exclusively so.

As you can see there, the oncology-side-specific databases and registries typically are populated by large-scale data that is not fully labeled, especially not to the extent to which we would need it to be in order to be able to classify something as large in dimensionality as the proteome is.

This is a slide where I'm discussing the data landscape hopefully in an illustrated format. On the right-hand side, you can see, so what we say, the external or large-scale data iceberg, which is short and fat. This is comprised of large-scale data that exists in registries and public databases. We're all familiar with these. This is extremely large data, but it is not very deep in terms of the number of features or labels that it carries.

On the left-hand side, we see something like the National Cancer Institute iceberg, which is tall and skinny. It has a lot of depth, including but not limited to the proteome as you can see there. These are sometimes relatively smaller datasets from studies, and these can be prospective or retrospective. Given the ability to treat patients on study, we're able to obtain biospecimens that can then allow for analysis of more deeper datasets, including the proteome, metabolome, lipidome, et cetera.

Growing data in this no man's line between the two icebergs is rapidly evolving. Our ability to be able to render these both structured, as well as analyzable, is highly important to be able to link it to the proteome, which at this point in time I would consider a more of a niche or a rare data type as it is growing. As it does so, the ability to place this in the public realm for people to work with and apply computational analysis to will be highly important.

My work specifically is the work of many people. I work with several stakeholders. You can see on the left-hand side there, including the American Association of Assistant Medicine with our effort of Operational Ontology Radiation Oncology or Oncology Now, O3, which I think will be published in *Radiation Oncology Journal* shortly. We also work with artificial intelligence resource here at NCI NIH, as well as Biomedical Translational Research Information System, or BTRIS, where we correlate electronic health record data together with other data channels, as well as computational genomics bioinformatics branch and computational systems biology branch.

All of our data at this point is integrated into NIH Integrated Data Analysis Platform, NIDAP, or Palantir for short. Today I'll be focusing more so on the proteome, given the time limitation.

In brief, here is the patient population that I will be discussing today and their proteomic analysis. These are 82 patients that have pathology-proven glioblastoma diagnosis, and their diagnosis date is between 2005 and 2013. This is an important parameter to keep in mind. These patients were treated with standard-of-care management, concurrent chemoradiation, and serum was obtained in these patients on trial prior to initiation of chemoradiation and after completion of chemoradiation.

The samples were analyzed subsequently using aptamer-based SomaScan proteomic assay. We have 7,289 human proteins. This analysis is based on 150 microliters of serum. The clinical data, which I spend quite a bit of time talking about, involves all of those aspects, including age, gender, how the patient was treated, radiation therapy parameters, as well as molecular parameters and outcomes, progression-free survival, overall survival, which were aggregated in NIDAP Palantir from my previous slide.

This is a good point to bring home the aspect of we spend 80% to 90% on the left-hand side of the slide, clinical data. Again, to my point of when we want to do computational analysis on oncologic data and aggregate that in a meaningful way with omic datasets, we need to keep in mind that the clinical data is both imperfect as well as highly heterogeneous, and it exists in multiple silos. It is often not aggregated.

The ability to collect and curate the data is paramount. Knowing that data and verifying its accuracy is absolutely the most important aspect of being able to create any sort of data linkage to be able to establish connections, as I described there in step three, and before we can aggregate data streams. Very often, the question that emerges is where is the PFS column, or where is the progression-free survival column? Very often it is expected that this column somehow magically exists in our datasets. It does not. Progression and the assignment of progression or coding progression is a huge problem in our field, which I'll come to in a second.

Now, proteomic data is very clean when we receive it. I think that there's always, or very often we are spoiled with normalization and a foregone approach to how this was obtained, including but not limited to perhaps the existence of where we have a pre-analytical testing that has occurred perhaps involving machine learning. I know SomaLogic has this, and we have carried this out on our data. Essentially then we go on to signal change and how this is being translated or linked to clinical data before we can get to machine learning.

Establishing that connection, I wanted to give an example. This is a relatively, I think, simple question to wrap our mind around, but a complicated question to actually answer. The question is, okay, but how much tumor does your patient have? The patient has glioblastoma, we know that, but how much glioblastoma does this person have? This is a question of tumor burden. Very often people say, "Well, just look at the resection status. Of course, that's logical because how much tumor was removed?"

That's not to say very much about how too much tumor is necessarily left behind and the person, given that I described this as a highly infiltrative tumor, and therefore that linkage cannot be immediately established.

Perhaps people will say, "Well, you design a target for radiation therapy, and those radiation therapy volumes could be used as a surrogate of how much tumor does the patient have." Of course, it is also highly imperfect. We do a little experiment here to illustrate this clinical data conundrum where we have these two parameters that have been shown to be associated with survival outcomes at prognostics or resection status where gross hormone resection involves removal of "all of the tumor." Subtotal resection is an incomplete resection of the tumor, as in the case that I described upfront in my talk.

Then we have some radiation therapy volumes. In this case, gross tumor volume, which is supposed to represent how much tumor the patient has, plus the resection cavity. This is our primary volume that we then grow to administer radiation therapy. Here I have actually contoured a residual tumor volume which is, in this case, exemplified by the tumor as enhancing one-to-one Gad or gadolinium MRI of the brain. What we can see here is that there is not great correlation between the type of resection status and the volume that emerges as a result of radiation therapy, dosimetry as rendered by variant eclipse. This is our treatment planning system.

The point that I'm trying to make here is that when we're asking the question “How much tumor does the patient have,” these surrogates are unlikely to be as helpful as we might think. Perhaps in the future, of course, we will look, and this is part of our ongoing analysis with artificial intelligence resource, is to look at using textural features and artificial intelligence through segmentation to pull out how much tumor may or abnormalities present in a patient's MRI that we can then link to the data.

As you can see here on the right-hand side, we know that even resection status isn't always prognostic in this case. For example, in this dataset, it is not. Then when we look at tumor volume as rendered from a clinical annotation of the scan used for radiation therapy planning in the lower right-hand corner panel, we can see that patients with a larger tumor volume, i.e. greater than four DCCs in fact do worse. That has been described in studies. It is important to know your data and keep these linkages that may be able to leverage in mind.

How do we establish connections to known molecular prognostic classification? I talked about MGMT earlier. This is O-6-methylguanine-DNA methyltransferase. This is a gene that encodes a DNA repair protein. By removing alkyl products from the O-6 position of guanine, what we are having here is that when the promoter is methylated. This renders it inactive, and therefore this results secondarily improved patient outcome.

Now, we know that MGMT status is prognostic and protective in glioblastoma. This is obtained from glioblastoma tissue. This brings me back that point I made earlier about how easy or difficult is it to obtain data parameters that can be linked to omic analysis.

If this is, for example, in our dataset, we know that methylated patients do better and unmethylated patients do worse, and unknown, of which there's roughly 37% of our cohort, we have an unknown MGMT status, for example, talk about missing this of data. Of course, I mentioned that these patients were treated 2003 and 2015, and this is not unusual to see in the literature. These aspects do need to be analyzed within datasets before we can move on to analysis of the protein or linkages.

We established connection to clinical parameters, as you can see here, both for several clinical parameters, as well as the outcome parameters of overall survival and progression-free survival. It is important to keep in mind that whatever biospecimen you acquire, in our case, it was serum for which we carried out the protein, which is what I'm discussing today, you want for those biospecimen acquisitions to be as close as possible to the intervention that you carried out or interventions, plural, in this case, because as the natural history of the tumor progresses, we know that additional treatments happen, perhaps different agents, different chemotherapies, repeat resection, reradiation, et cetera. All of these are likely to alter what you're going to see when you do an analysis of the, for example, proteome.

This brings me the proteomic signal. We assess for alteration between time points, in this case, prior to administration of chemoradiation and after chemoradiation. It is important to be able to look at this logically in terms of significantly altered targets, significant patient numbers and significant target interactions before we can look at hallmarks of cancer and hypothesis testing. In this analysis, in this cohort that I described earlier, this is just to give you a snapshot, this is from a recent publication of ours in Cancers, a snapshot of the top 10 upregulated and downregulated proteins. I knew that there was signal in the data before I moved on to more complex analysis, which I will show you shortly.

Given that, when we did then more or several bioinformatics analysis for signal alterations, we again noted that there was a signal alteration between pre-chemoradiation versus post-chemoradiation in the proteome. We also observed some really large changes pre- versus post-chemoradiation for some subjects. This prompted a lot of discussions both within bioinformatics within our teams here, as well as with SomaLogic. We did carry out several additional analyses looking at outliers. Ultimately, we kept all of the data. We did not exclude our outliers, and I can discuss this in more depth later, if questions arise.

Very important, the serum samples I explained were obtained over a large period of time, and they were kept in the freezer for an average of 3,442 days. There's quite a range there as you can see from 2.2 to 15.9 years. I did not find an alteration in signal, given the storage of these samples. I'm just giving three examples of three targets there, so that you can have a look at comparison of how long the sample was in the freezer and subsequently the protein signal change in blue versus the days collection to analysis in yellow. This was good.

This is not an assumption, by the way. It may be that given the type of analysis you carry out or the type of protein that you look at, you may or may not find a difference.

We did not, however, find a difference based on how long the sample was in the freezer. Now, interesting, in terms of looking at signal alteration, which I alluded to earlier, pre- versus post-chemoradiation, I took a look, initially, at several factors, and this is just one example. I looked at long-term survivors. These are patients that survived more than 20 months. I will just quickly flip my slide and go back and forth to give you a context for this versus short-term survivors.

When I go back and forth, you can appreciate in this panel that there is a significant just a glance alterations that you are able to see in some of these proteins. Here we selected some that were either extremely upregulated or downregulated and proteins that they interacted with, so that you can get a sense of that alteration signal. The alterations are in aspects that are both interesting, as well as a hypothesis-generating, in the sense that there's tumor suppressors and there are angiogenesis mediators, hypoxia mediators. Some aspects that have yet to be defined in more depth. With that, I go on to even another analysis where we looked at rapid progressors.

These are folks that progressed in less than six months following the diagnosis versus long-term survivors that survived more than 30 months. Again, you can appreciate that there is an alteration in signal in some of these important domains. With that, we arrive as if we are to take these signals, and I cannot show you all of them, because this data is embargoed at this moment, given that a couple of our papers are sitting with reviewers or conferences. We arrive at this problem here.

When we take our signals, we arrive at the hairball problem. The hairball problem is basically this conglomeration of signals that we really don't know how to take this into bench to bedside and what signal to go after, given that we don't quite know what is meaningful here, what is relevant, and what is irrelevant. How do we look at removing relevant or insignificant findings? Which brings me also eventually to computational analysis. Really, the point here is for ongoing work. Just to give another context, this is the administration of HDACi inhibitor, valproic acid on study, which was the case in 29 of these 82 patients.

You can see how there is this extreme explosion, this firework of signal that we're seeing. When you superimpose the full change, you can see all these targets. Some are going up, some are going down. What do we do? Well, we would like to be harnessing artificial intelligence. I have spent most of this talk discussing processing aggregation of data endpoints, as well as how to plan for transferability and validation in the future by linking clinical data to proteomic data. With that, I think this is the final crowning glory here, is harnessing artificial intelligence in the proteome.

The only way that is possible is if all these steps have occurred prior to that. When your computational analysis folks show up in the lab or in the clinic and want this perfect data set, it is important to note that it is actually a lot of emotional labor to arrive at that point before you can harness AI. That's the process that I described, and now here you are. Now you have the proteome, and that is amazing. And you want to combine that with other features, including imaging and radiation therapy. That is evolving work for us, but we did look at feature engineering to combine clinical data and the proteome as acquired, based on my previous slides.

We want to look at the dataset, create classes, and determine what matters most. For that, we use artificial intelligence-based approaches for prediction. This paper has just been published in the journal *Cancers*, I would encourage you to take a look at that to look at our-- this is the work of Dr. Toshi, who's a post-doc in my lab. You can look at all of the development and the experimentation that went on to be able to arrive at this rank-based, hybrid feature and waiting selection method, which I will not discuss in great detail here, because it is described over the course of 27 pages in this manuscript.

Given that we were able to identify several relevant proteins, of which six of them had emerged in all of our other analyses pretty well throughout. That's important to see, that we were able to identify similar signals, and this was carried out by several groups in parallel analysis. We also identified the cystatin E/M or CST6, which is a cysteine protease inhibitor. I'm only pointing this one out, because this was not one of our highly emerging signals initially in any of our other analyses. But when we carried out this analysis using this machine learning approach, this one emerged as one of the seven.

These are the seven relevant proteins that define the administration of chemoradiation. When we look at signal proteome acquired prior to chemoradiation and after, these are the seven that emerge, that define that interventions given that there are two that are given concurrently. We can see from this that these proteins are actually highly relevant in glioblastoma. And I don't want to say just in oncology, but in glioblastoma, that is highly important. To carry out some mental tests in terms of what evidence exists out there, because this will not only allow for potential validation down the line, but also to make sure that linkages can be obtained in terms of signaling pathways.

Which brings me to my next slide. When we now take these seven features that emerged via this method and we place them here, in this case, in ingenuity pathway analysis, we now arrive at these disease and functions. We see, for example, glucocorticoid receptor signaling. This is interesting, because of course patients do receive dexamethasone or other corticosteroids during the administration of chemoradiation, prior and during, given that they have very often significant swelling and inflammation both due to the tumor itself, as well as the administration of treatment.

That particular aspect hasn't really been classified and is a very difficult problem to classify in and of itself. We also found PI3K/AKT and 14-3-3 mediated signaling that emerged as part of disease and functions. These were part top canonical pathways that these seven features correlated with, as you can see here. This is important, because we know that P13K/AKT in glioblastoma is associated with development and progression by growth and survival metabolism and resistance to cell death and angiogenesis.

This is important, because we have seen these signals in several analyses. What is really fascinating is that we can see some of these signals be both up and down, but in different patient populations, which gets me to my point upfront about the biological heterogeneity and how the proteome may be able to give us a peek under the hood as to what actually happens when some patients respond, and some have disease that seemingly is extremely resistant and has ongoing progression.

We go from this, our hairball problem, to identifying upstream mediators, in this case, EGF and CTNNB1, as well as going from this explosion of signal that we saw in the proteome pre- versus post-chemoradiation in glioblastoma to a more cleaner pathway. This may be what happens proteomically when you give an individual with glioblastoma chemoradiation pre- post-analysis of the proteome. The signals that you can see there have the **[unintelligible 00:40:18]** change-- the proteins rather, have **[unintelligible 00:40:20]** change superimposed with the data from our panel.

You can appreciate that some targets have predicted activation, and some have predicted inhibition. All of this is subject to ongoing research under investigation as we explore the balance between proliferation and survival and tumor migration in the microenvironment. We come to part four and the final component of my talk regarding current intersectional research questions.

I started with this slide earlier on to give you some context as to if you were to find data in this space, where would it exist? We know there's tumor tissue. We know there's some laboratory values that are standard of care. And we know that there may be later tissue once the patient has progressed, but this is harder to come by, really, and not as common, given the fact that we are discussing tissue from the brain.

We know that MRI is one aspect, so imaging that we have the most of. Given that I just explained that there is signal alteration in the proteome in this context, what would happen if we had serum proteome, for example, at all of these time points in larger data sets? Imagine the ability to be able to analyze biological signaling pathways in that context. Imagine if that data was publicly available for researchers to explore and be able to train and test artificial intelligence algorithms. This needs to be done with clinical data interplay and clinician supervision to be able to arrive at interpretable findings.

I early on planted the thought about how we don't fully know what progression is in this space, how treatment effect and tumor progression are difficult to separate when looking at imaging. This is the subject of ongoing work with artificial intelligence resource here at NCI NIH for our data set, which will then be placed together with omic analysis to further examine the biological interplays of signals that I described earlier. Defining progression is a very challenging problem. And although there's been significant evolution here, we have yet to arrive at an artificial intelligence-based response assessment criteria in neuro-oncology.

This is really important, because as you can see in this very busy table, when I have looked even more recently at response criteria, this is a complicated problem where we have tried to quantify what response or progression is. But you can see it's highly imprecise, and this remains the case even today. It is possible that by harnessing omic analysis, including but not limited to the proteome, we may be able to define this in a manner that is robust and then can define the progression time point as or outcome endpoint to be able to take a more robust or a ground truth sense of what progression is into artificial intelligence analysis that harnesses the proteome to be able to peel apart that onion of heterogeneity and redundancy.

It is very important to be able to place radiation therapy dosimetry data in clinical frameworks so that it can be used. It is currently in the treatment planning system Varian eclipse. While it is available, it is very difficult to aggregate data to be able to take all of these data streams together to be able to take advantage of computational analysis.

We have done so here in the NIH data integration platform **[unintelligible 00:43:54]** Future directions involving these data sets, I think, involves the aggregation of data that exists in all of these channels, and being able to bring it today together in a robust defined manner, both rendering data that is currently unstructured data analyzable.

This is absolutely crucial to be able to harness computational AI-based method. Simply having the method is insufficient to be able to advance the brief if we have data where it's, as you've heard the term, garbage in, garbage out. It is that effort that is occupying a great proportion of the time where probably computational-based methods also need to be employed.

The proteome, of course, is a very exciting but deep component of the data that needs exposure and aggregation. This is part of the Cancer Moonshot goals, in terms of creating new data systems that break down silos and ensuring that knowledge is disseminated, made available to as many people studying the subject as possible, so we can make a difference for patient outcomes.

In summary, I discussed clinical data exposure. I discussed some proteomic data in our glioblastoma patients that was obtained using serum pre- and post-chemoradiation, and the usage or example of computational analysis to be able to pull out the most significant features. It is my hope that in the future we can carry complete data sets into the public repositories as the data continues to be published and is increasingly peer reviewed.

With that, I would like to say thank you to all of my collaborators and my staff, especially my core team and radiation oncology, and Dr. Tasci **[unintelligible 00:45:44]** Sarisha, my student at this time, as well as all my other students, radiation oncology branch, and all of my collaborators in computational **[unintelligible 00:45:53]** bioinformatics branch and computational statistics, as well as artificial intelligence resource. With that, I'll stop there, and we'll take any questions.

Rebecca: All right. Wonderful. We can move on now to the Q&A. As a reminder, you can still submit questions using the Q&A tab on the left of your screen. I see we have a lot of great questions already. We'll try and get to as many of them as possible. Let's begin. What led you to begin exploring proteomics in GBM, and what specific strength/utility does it provide as both a single omics approach and in combination with other multi omics?

Dr. Andra: I think this is a great question, because this subject comes up a lot at a lot of conferences and a lot of papers and abstracts as I showed a bit in my presentation there. The reason that this was on our radar for a long time, we collect biospecimen on protocols as you saw there and have been doing so for quite some time. The goal of that biospecimen collection has always been the analysis, in this case, both blood and urine. You guys saw serum proteomic data today, but we are looking at additional analysis, including involving **[unintelligible 00:47:12]** et cetera.

These ideas, given the fact that we have so little understanding of the heterogeneity and redundancy that occurs in a signaling standpoint in glioblastoma, we wanted to be able to probe that. We knew that this was only going to be possible through biospecimen acquisition. The proteome on this scale, the 7K scale, only became available more recently. We knew that eventually this was going to be something that we wanted to do, hence our biospecimen acquisition and the intent to carry out this type of analysis.

To answer the question, I think that we need to look ahead where we think we will be in perhaps 10, 15, 20 years and plan accordingly in our prospective studies. Because it is highly important, as I cannot stress enough, that we gather those biospecimens that are realistically analyzable. Tissue is a problem, of course, when we're talking brain. We want to be able to have biospecimens that have the potential to be interrogated omically down the line with multiple time points. For us, the proteome was on our radar, hence our biospecimen acquisition.

Just to touch on the single omics approach in combination, I haven't specifically talked about the concept of analyzing the pet protein. I refer to the pet proteins where we analyze 1 protein or 5 or 10 or 20, but not 7,289. My issue with that approach I think is that as I try to show the redundancy, you may be missing all of these interplay, this balance of signals that you're not really going to see if you analyze five proteins or five-protein panel.

I showed you there 14-3-3 sigma as a potential mediator, but if I only looked at 14-3-3 sigma, for one thing, I would not know what 14-3-3 sigma is talking to. But I happen to know that it has another 6 friends in the panel, for example, as well as another 100 that are talking to it that modulate its activity. I would not know that if I was looking at an approach that had a more limited breadth of proteomic analysis, if that makes sense.

Rebecca: All right. Great. Thank you so much. That's a really extensive answer. Our next question, in your experience, what are the strengths of the platform for use in both translational mouse models and downstream clinical research?

Dr. Andra: We have considered both of those aspects along the way while we carried out this analysis. We think that the strength of the platform, which I alluded to already, is that it's very large. I think to me that's really important. That's important also in connecting it to mass spec data sets or other proteomic data sets that are out there, including but not limited to, I only talked about serum today, but there's of course tissue proteomics, et cetera. The more signal you have to work with, that you can link to other signal, the more you're likely to find a meaningful linkage that can help you eventually identify druggable targets to improve outcomes.

I think that bench to bedside component that you alluded to is achievable given that. That same connection needs to be built to animal work. Even though the measurements are not comparable head to head. These are discussions that we've actually had with SomaLogic several times. We know that there are protein panels that are available in animals, and of course, a lot of the data will be originating there. I alluded to that in my talk given what we see in web of science in terms of existing data.

We know that that linkage can be built between human proteomics, specifically GBM proteomics and animal GBM proteomics. That linkage, once built, which is also going to be computational, is going to allow us to go back and forth between the findings to be able to dig deeper and to be able to manipulate the pathways with drugs, et cetera, to be able to figure out how outcomes can be improved.

Rebecca: All right. Great. Our next question, what's next for your research and proteomics and or multiomics applications in these areas of high unmet clinical need?

Dr. Andra: We will continue proteomic analysis. Actually, we are going to do more samples. We're going to analyze even more samples. This iterative process. You asked me how did we start doing the proteome, and I said that we had a bit of a vision for our program and what we wanted to do. We continue to acquire specimens on study, and we are going to carry out more proteomic analysis on additional specimens. We will also carry out metabolomic analysis. We are looking at other options as well, including linking this with the tissue that we have for these patients. This is often challenging in the neuro-oncology field, because we don't have a ton of tissue, and everybody wants the tissue.

It is a problem in terms of being able to not use up too much of it, but also to be able to link what we see at the genomic and transcriptomic level with what we see at the proteomic level. That's our four-pronged approach. Metabolomic, further proteomic, linkage to tissue, and linkage to imaging. I didn't have the time to discuss all of that today, but we are using computational analysis and linkage between all of these aspects.

Rebecca: All right. Thank you. Yes, sorry, I'm just scrolling a little here. Next question, what are a few of the most exciting or surprising findings and insights from applying an unbiased proteomic approach here and in combination with other high-resolution methods versus pursuing hypothesis-driven biomarkers?

Dr. Andra: I think the most surprising-- I think the first thing was that we have signal, actually, and this sounds-- This is an interesting question, is that I think there should be no **[unintelligible 00:53:53]** assumption that when you take a sample pre- versus post- an intervention, that you necessarily have a change in your signal. I don't think that should be an assumption, but there is signal, and there's a lot of signal. In fact, there's fireworks of signal.

I think that was the one aspect that at first we're like, "I wonder, these are proteins. This is the proteome." 7,289 signals. Some of the changes are relatively small. We know that the proteome is 20,000 some proteins, of which we believe about 600 or something like that are functional. Really, there's no assumption that you're going to see signal, but we do see signal.

Coming to the next surprising approach-- We can argue about how surprising this is or isn't, but the administration of HDACi inhibitor, valproic acid, as an agent together with chemoradiation on study, looking at that aspect was interesting. We saw a signal there, too, between the folks that had received the agent together with chemoradiation and those that had not, that had received chemoradiation alone.

Then the question is, well, when we see signal situations like that, how does that gel or not gel with folks that, say, receive this particular agent for seizures but don't have a brain tumor? That is an interesting aspect, as this gets me a little bit up to the point of how are we measuring or linking this to the tumor state or the tumor burden? Those were surprising aspects.

Surprising also, but in a sense reassuring, was to see that we were seeing this balance of signal between aspects that were elevated or proteomic panels or groups or pathways that were elevated but also decreased. It was this constant balance that we were seeing in our data that we still see and to be able to then look at how this gels with clinical data in terms of outcomes, because this indicates that we are looking at a biological interplay of multiple signals.

Some go up, some go down, and when we put this in, say, ingenuity pathway analysis or **[unintelligible 00:55:55]** GSEA, we really do see that some things go up and some things go down, but it's really the composite of that entire package that you're measuring. That was also interesting. Maybe it would've been a bit interesting to see that only some things only go up and some things only go down, but it's actually an interplay. It's a balance. That is really important aspect to keep in mind. The use of artificial intelligence when we arrived at these seven features that you saw there, and this is an analysis where we really only looked at pre versus post, what seven things emerge, in this case, seven things.

It was interesting to see that those seven proteins did link to biology, did link to glioblastoma in a way that could be conceived of as biologically meaningful and jelling with evidence-based medicine as currently captured. As data evolves, I'm sure that all these findings of course continue to evolve, but that was really interesting. To find one of the signals wasn't a really prominent one, but did emerge in the machine learning analysis. I think that's an important aspect actually, because it allows us to remove irrelevant or insignificant features to elicit a high correlation. Because people might say, "Well, why didn't you just take the top 10 or top 20 or top 100, 200, 500? How would you know which ones to use?"

You wouldn't know. In the non-ML analysis, we found anywhere from 200 to nearly 400 proteins that "mattered." When we did the AI machine learning analysis, we found seven that defined this particular aspect. That's pretty cool.

Rebecca: Well, great. This I think will have to be our final question. We are coming to the end of our time here. We have had a lot of great questions, and we couldn't get to all of them. But we'll try and get back to everyone personally after the webinar. For our last question, thank you for the excellent talk. You mentioned the presence of outliers. Interested if you might expand on this and the degree of individual variation or subtypes observed in these patients. Have you detected what clinical features this may be associated with?

Dr. Andra: Great question. This was a subject of discussion for weeks. We really did not find a specific correlation with any clinical features. I talked a little bit about how our first thought was, "Okay, some of these samples are quite old. They have been in a freezer a long time. What does this do or not do to the signal? We don't actually know." We didn't find a correlation there. SomaLogic actually helped us with this and carried out an ML-based analysis, which I didn't get a chance to really even discuss in great depth here. This is something that is available to people. I'm sure there's also individualized other analyses that could be carried out that are perhaps SomaLogic independent.

In any case, that didn't turn up anything. When we looked at all of the parameters in terms of validity of our data, there was no good way to say that this is either linked to any clinical aspect or any storage aspect. Ultimately after discussion between all of the groups that I mentioned that are my collaborators, as well as SomaLogic and our team, the decision was made to keep all of the data, because we were concerned that by not doing so, we would be introducing bias.

Bias is one of our big things that I also didn't get to discuss today, but we really were very conscious of the fact that we wanted to be as true as possible to the signal as captured. It is possible that there are individuals out there that are just natural outliers. There may be features that we do not know or cannot define at this point given available clinical data, because remember, the clinical data I use 29 features, but you cannot measure the person ran to the laboratory that morning where they had a particularly greasy meal or whatever the situation might be. You don't know that. There's no way to actually quantify that aspect. For now, we kept all of the data. We didn't exclude anything.

Rebecca: Well, thank you both so much. I believe we are out of time now. Thank all of you for attending today's Fierce Biotech Webinar, for submitting these great questions. I'd like to thank our speakers for participating, and SomaLogic for presenting today's webinar. A recorded version of this will be available for you to access within 24 hours using the same audience link that was sent to you earlier. Thank you again for joining, and we look forward to seeing you at future events.