Transcript

***Disclaimer:** This transcript has been generated using AI technology and lightly edited. It is intended to provide an accurate, verbatim representation of the language used by the speakers during the RECOVER Research Review (R3) seminar. Please note that some errors or omissions may have occurred due to the limitations of automated transcription. Videos for this and previous seminars are available on the RECOVER R3 Seminar Series web page.*

Patrick Ahearn:

So good morning and good afternoon, everyone. Welcome to today's webinar. I'm going to pause for just about 15 or 20 seconds to allow everyone to get connected to the audio, but we will be getting started momentarily. Thank you. So once again, welcome everyone to today's RECOVER webinar. This is Part 1, "Using RECOVER data to advance your research." My name is Patrick Ahearn with RTI, and I'll be helping out with the virtual room today. So just a few quick housekeeping reminders before we get started. When you log in today, your microphone and camera will be automatically muted. But if you have any questions for our presenters today, please submit those in the Q&A window at any time. And if you run into any technical issues, please let me know in the Q&A window as well. Closed captions are available during today's webinar, so just click on the show captions button on your main Zoom toolbar to turn those on. So, at this point, I'll go ahead and turn things over to Christine to kick off today's session. Thank you.

Dr. Christine Bevc:

All right. Thanks, Patrick. And welcome everyone to today's R3 seminar. My name's Christine Bevc and I'll be your moderator. Next slide, Patrick. As we mentioned, today's session kicks off our 2-part series focused on available RECOVER data and biospecimens. We'll be covering a high-level overview of the observational adult and pediatric cohort studies with a quick refresh of the study design, the participants, and the topics captured in RECOVER's large nationally representative dataset. You'll learn where to find the detailed study documents and understand how the data is structured, as well as how to access and analyze the data using tools with NHLBI's BioData Catalyst. Biospecimens is going to be covered in Part 2, so we'll see you next month for that. But today's goal is really to help you understand the data and how to get started. All right. Next slide.

The data discussed today during the session represents the contributions of thousands of RECOVER study participants since we've been collecting data since the observational studies began in 2021. And respect for and the protection of the interests of research participants is fundamental to NIH's stewardship of people's personal health data. We encourage you to review the RECOVER data access and availability guide that was distributed to registered participants in advance of this seminar. And I think Kate is going to share that also in the chat as well. But this is in compliance with NIH's data management and sharing policy. RECOVER allows those who are interested to view the data and researchers also to be able to analyze it, learn from it, and incorporate it into future studies. Next slide. One of the ways that researchers can apply to use RECOVER participant data is through our current funding opportunities. This series directly aligns to the current Research Opportunity Announcement and the NIH's Notice for Ancillary Studies.

In addition to ongoing NIH extramural funding, researchers both within and outside of RECOVER can apply to this opportunity to use RECOVER data and biosamples in their proposed research studies.

The link to this opportunity is, I think Kate's also going to drop that into the post as well, but you can learn more and by either clicking on that link or going to recovercovid.org/funding.

All right. So please join me in welcoming our panel of experts for today's seminar. They bring a vast wealth of knowledge with them. So, it's my pleasure to welcome back Dr. Leora Horwitz and Dr. Rachel Gross. These are 2 of the incredible principal investigators from the Clinical Science Core at NYU Langone. Dr. Horwitz directs the adult observational cohort study, and Dr. Gross leads the pediatric observational cohort Study. We're pleased to also welcome James Chan and Trisha Balan from the RECOVER Data Resource Core at Massachusetts General Hospital. James Chan is the lead biostatistician for RECOVER, and Trisha Balan is an applications analyst, and together they're going to be discussing the data, tools, and skills for accessing and analyzing all the RECOVER study data that's collected from participants. And rounding out our panel, we're joined by Emily Hughes and Cera Fisher. Ms. Hughes is the product manager for BioData Catalyst powered by PIC-SURE, and she'll be introducing us to one of RECOVER's data exploration tools.

Ms. Fisher is a program manager for NHLBI programs at Velsera and supports BioData Catalyst powered by Seven Bridges. Now, this panel has an incredible amount of information to share with you today, and I want to start by thanking everyone who submitted questions in advance. And a quick reminder, as Patrick mentioned, you can submit your questions at any time during today's presentation using the Q&A feature in your Zoom menu. After today's panel, it's going to be a full day. Our speakers are going to try and answer as many questions as possible. But please note, we can't answer any questions about individual cases or analyses, but we'll make sure that we can get to as many of those as possible. And with that, we are going to jump right in and I'm going to turn things over to our first speaker.

Dr. Leora Horwitz:

Thank you, Christine. I'm Leora Horwitz. As she mentioned, I'm responsible for the adult cohort here at the Clinical Science Core at NYU. If you want to use data, you need to understand how the studies work. So, I'm going to just give a brief overview of how the adult observational cohort study is set up and the sorts of data that we collect. You can go to the next slide. As with both of the large observational cohort studies in RECOVER, we are asking some core questions. How many people get Long COVID? Why do some people get Long COVID, and others do not? What are the risks for Long COVID? What symptoms do people feel when they get Long COVID? How can we, in other words, identify or define it? How long do people feel sick when they get Long COVID? What's the trajectory? And finally, what causes Long COVID? What goes on in the body? What are the pathophysiologic changes that are happening in people with Long COVID? Next.

The adult study is being run at 83 sites in 33 states across the country, plus Washington, DC, and Puerto Rico. This is highly geographically diverse. Next. The study design allows people to enroll in the study at any time and at any state of infection. People can enroll whether they have had infection and whether they have not. People can enroll within 30 days of their very first infection. Those are these people in the blue boxes on the top left. They are then followed forward every 3 months for the rest of the study. People can also enroll after they have had their first infection at any time—months, years, even—after their first infection. Those are the people in the orange boxes, square boxes. They too are then followed forward. People who are uninfected again just enroll. And of course there is no time since infection for them.

So, we have all different sorts of populations in this study, people who are infected in the very earliest days of the pandemic, and people who were infected much later. Next. Altogether, we've enrolled over 14,700 participants in the adult cohort, including over 2,000 people who were pregnant at the time of their first COVID infection. Almost 5,000 of these were enrolled within 30 days of their first infection and then followed forward. Over 7,000 were enrolled more than 30 days after their first infection. And over 2,000, 2,400 folks were enrolled who believed that they did not have any COVID infection at the time that they enrolled. Since enrollment, of course, well over 1,000 of those participants have gone on to have their first infection, but at the time that they enrolled, they did not. Next slide. Our participants look very much like the US population. As a whole, our distribution by race and ethnicity is quite similar to the US average, with the exception of an over-enrollment of the Asian population.

Next slide. Our population is, however, over-enrolled for female participants. It's about 60% female, about 40% male. Next. So, the way that the RECOVER adult cohort works, again, it's recruiting from around the US with and without COVID, pregnant or not pregnant at the time of initial infection. And everybody who is enrolled in the adult cohort undergoes some basic activities that I'm going to tell you more about in a minute. These basic activities include questionnaires, surveys, laboratory studies, biospecimens, and a minimal clinical exam. People who have certain symptoms and random assignments of people go on to have additional testing. It's both low risk clinical tests that we call Tier 2 and more invasive or more complicated tests that we call Tier 3. Let me explain more about that. Next slide. So, the study schedule looks like this. There are activities every 3 months for participants in the study.

Most of those activities are surveys. You can see those dots in the columns going across every 3 months. All of those are surveys. At the bottom of this graphic here in the orange, you will see that office visits, in-person visits, have been at enrollment, of course, at about 6 months after the initial infection, if the person enrolls before that, and then yearly thereafter. And the same is true for biospecimens and lab studies with the exception that we do also collect biospecimens and clinical labs at 3 months after infection. People who enroll a year after their infection or 2 years after the infection skip those activities, and they just carry on the schedule according to the time period since their initial infection. Or if they're uninfected since a negative test. So, you can see that we have in-person assessments roughly yearly, including biospecimens and laboratory assessments.

Next. Our surveys are very elaborate. In fact, they take our participants many hours to complete, and yet they complete them with an extremely high level of completeness. We are missing very little data, in fact. Our survey topics cover basic demographics, age, sex, race, ZIP code, baseline disability status, vision, hearing, physical function, and so on before the COVID infection, baseline comorbidities, in other words, diagnoses that people had before their initial infection. And then those same conditions we keep track of if they newly develop during the course of the study. We collect details of each COVID infection, including how it was diagnosed, how it was treated, and when it occurred. We similarly ask people if they are receiving treatment for Long COVID symptoms. We ask about all the medications that people take, including supplements. We ask about any vaccines they've received for COVID, including the type of vaccine and the date of the vaccination.

We ask about alcohol and other substance use. We ask a large number of questions about social factors, some only on enrollment and some every 3 months. We ask about marital status, gender, and sexual identity, employment status, income, education, language, insurance, all kinds of other questions about the environment in which they are living. For those who have had a pregnancy at any time during the study, whether it was during COVID or not, we ask about the outcomes of that pregnancy. And then we ask a lot of questions about self-reported outcomes. We ask about global health, mental health, including depression, anxiety, PTSD, grief, and then we ask about Long COVID symptoms. On the next slide, you'll see a list of all of the symptoms that we ask about. Well over 40 different types of symptoms, covering just about every organ system, because Long COVID is a multi-organ system condition.

So, we ask about general sorts of symptoms. We ask about problems with the heart, with lungs, with brain, intestines, reproductive symptoms, and so on and so forth. Wide variety. Next slide. The Tier 1 that we talked about, the things that everybody does, includes not just those questionnaires, but also some basic examinations. We collect vital signs, waist circumference, 30 seconds sit to stand, which is a measure of how easily people can stand up without using their arms and then sit down again, repeated over 30 seconds, good measure of frailty and physical function. And we measure the 10-minute active stand test, which is a measurement of blood pressure and heart rate while lying down and also while standing up for 10 minutes to assess autonomic dysfunction. Then we collect a wide variety of basic clinical laboratory tests, which you can see in the table below, as well as specimens for storage at Mayo for use in future research. And we collect blood, plasma, serum, PBMCs (peripheral blood mononuclear cells), and so on, saliva, urine, stool, and nasal swabs for future collection.

People who enroll at the time of enrollment who tell us that they did not have an infection, we also test antibodies in those participants to make sure that they did not have an infection they didn't know about. Next. For people who have certain symptoms and also a random sampling of people, we assign additional testing. We call those Tier 2 and Tier 3 tests, and those are more complicated sorts of tests. So, here's an example of how we assign, for example, here the 6-minute walk test.

People who have a Modified Medical Research Council score of one, in other words, shortness of breath, people who tell us that they're very fatigued, people who have cough, people who have post-exertional malaise, people who have low oxygen in their blood and so on. These are all people who, if they report those things, are offered the opportunity to take a 6-minute walk test. In addition, we offer a random sampling of people who don't have these conditions, an opportunity to take a 6-minute walk test. So, on the next slide, you can see a list of all the Tier 2 tests that are offered to people with relevant symptoms of their own.

Of course, the symptoms for 6-minute walk are different for the symptoms for, for example, a smell test, but each of them is offered to people who have symptoms that might be understandable with those tests. We collect laboratory tests; we do clinical assessments. As you can see here, we collect radiology studies and other procedures such as pulmonary function testing. And again, a random group of people who don't have the relevant symptoms are also assigned or asked to take these tests.

Next slide. People who have symptoms also may be offered the opportunity to do more invasive testing. We call these Tier 3. We offer these to fewer people because they're more invasive, more expensive, have a higher risk of complications, but these include things like hearing testing, complete neurocognitive testing, lumbar puncture, colonoscopies with biopsies, skin biopsies, and so on. And again, these are offered to people with symptoms. In some cases, we also offer them to a random sampling of people, but that depends on the severity of the sort of invasiveness of the test and the degree to which we feel it's safe to do so. That completes my overview of the adult protocol.

Dr. Christine Bevc:

All right. Thanks, Leora. Next, Trisha and James are going to dive into the RECOVER data and structure and how all of this information is organized for the adult observational cohort. So, Trisha, go ahead and get us started. Let's dive in.

Trisha Balan:

Hello, everyone. My name is Trisha Balan, and I'm an applications analyst at Mass General Hospital. I work on the RECOVER project as part of the DRC, along with our lead biostatistician, James Chan. Today, we'll be giving you a brief overview of the platforms that we use at the DRC and share a bit of a snapshot of the data collected as part of the RECOVER study. So first off, REDCap is our primary data collection system for RECOVER and houses all the ECRFs, electronic case report forms used by study coordinators to enter visit data, chart reviews, and test results. Participants also complete surveys through personalized links. Next slide, please. REDCap also contains the data dictionary, which contains variable names and what information they collect. The data dictionary within REDCap is essential. It defines every variable collected across sites and ensures consistency as data is shared. Next slide, please.

So, after data is collected in REDCap, it is shared with the broader community through Seven Bridges. The secure platform gives investigators, biostatisticians, study sites, collaborators, and stakeholders access to RECOVER data for analyses. Next slide, please. And the DRC primarily uses RStudio as a main platform for analyzing RECOVER data. Our current analysis scripts and workflows are written in R. We use R for data cleaning, quality checks, visualization, statistical modeling. Working within a shared platform gives us and helps us maintain reproducible workflows and collaborate efficiently across our different teams. Next slide, please.

So, if you're an investigator who would like to work with RECOVER data, we recommend reviewing the network of biostatisticians, NBR training materials found on the NBR website linked here. Because each cohort has a complex protocol and structure, these courses help orient researchers to our very big dataset. These materials provide a good foundation, although most analyses will still benefit from the support of experienced programming and statistical staff. So, with that quick overview of core platforms, I'll now hand things over to James. We'll walk through a bit of the adult cohort data.

James Chan:

Thank you, Trisha. And I'll emphasize that the NBR materials are a really great overview of everything that's in the data. There's a lot of details in there, too much to really cover in this kind of talk. So, I do point everyone towards the NBR materials that are really excellently put together. I am going to give what's really an example of some of the data that is in our system to help people understand where their numbers might end up for any analysis that they're planning. And we have been, earlier on, ROA was mentioned, we've been working with ROA investigators for the last few years, and the examples here are reflecting the issues that came up for them and the things that best help them understand what the data looks like in our study. So, starting with this slide, I'm using the symptoms form mostly here as an example.

The availability of the symptoms form is a pretty close reflection of the other survey data in our study, and often it's the most important survey. So, what we're seeing here is how many forms are available at each, what I'm calling here, a visit month, which is really just a visit block or a month period after the initial infection for a participant. Because it's every 3 months survey, our study splits these visits into 3-month blocks, and this reflects what Leora was describing earlier as participants who enroll acutely in their initial infection. And then the participants who enroll months or years later. And you can see here, if you have an analysis that needs the 24-month symptom data, you're in the ideal place for how many forms are going to be available for that analysis. If you wanted the symptom data from right at the first COVID infection, you have much lower numbers because many people enrolled months after that, and of course didn't fill in a survey form at that time.

And there is some retrospective data collected, but as far as prospective data, they're not going to have a form at that time. And similarly, if you're looking for something several years out, you're going to see a taper of the data because many people just didn't make it to that length of follow-up during the study. We can go to the next slide. This is the same data. So, we're looking at the exact same thing. All the numbers are the same on these bars. I've just overlaid the sample collections. So here, hopefully it's clear to see the symptom forms are collected every 3 months, and the sample collections are on a different schedule. For acute participants, they have 3 sample collections in the first 6 months. Everyone else does annual collections after that. So, emphasizing here, if you are doing an analysis that only requires symptom forms, you're going to have more data than if you're using analysis that does require some samples.

We can go on to the next one. This is the same data as we're looking at before, but for people without a history of COVID infection, we have much less of these participants, and we see a stronger taper over time because people move from the uninfected category to the infected category as they get infected on study. And we can go to the next slide, and this is a similar overlay here of the biospecimens forms. The participants without a history of COVID infection do follow the same schedule as everybody else. So, it is one schedule for all the participants in the study. So, you see the same every 3-month forms and annual biospecimen collections.

And I should say here, I'm using the same 3-month visits. However, it is just really about the setup of the study where we assigned a pseudo-infection date or an index date for all the uninfected participants as well that aligns with a negative test. So, this does help show what the schedule of visits are, but depending on what your actual question is, it may not matter at that time point. Okay, we can go to the next slide. Just a quick note that we do have longitudinal data. I just want to emphasize that here. That is what we were looking at in the previous slides as well.

Many of our participants over the last 4 years have entered more than 10 visits. So, we have many, many symptom forms and other forms for those participants and several sample collections as well. This is showing the numbers we have in the infected participants. And the next slide, if we can move on to that, will show the same thing for the uninfected participants. And again, the taper here is partially due to the same data structure as the infected participants, but also those people crossing over from uninfected to infected. And so, we can go to the next slide.

So, I'm showing here the same slide that Leora was showing before. This is our schedule of assessments. I'm going to use this to just talk through some high-level counts for some example analyses just to show how numbers might change as you're thinking through what your actual analysis is. So, we can go to the next slide.

So very high level here, we have about 15,000 participants. They've contributed about 140,000 symptoms forms. That's about 9 per person. Some will have more, some will have less, of course. And about 50,000 visits with a sample collection. Now, there are many different types of samples. So, 50,000 visits may have had a sample collection, but exactly what was collected at a given visit varies. So, we can go on to the next slide for an example.

Okay. So here is a very simple analysis question that I have as an example. So, you have a serum assay you want to run on serum collected within 30 days of first infection, and you want to compare that to symptoms reported 12 months afterwards. So, the first step here is to look at that sample collection. The 30 days doesn't quite align with the visit schedule we have, but we have a collection date, so we look at that 30-day period. We can narrow down to just serum. And for the samples available, we have 3,300 participants with that sample available. As we narrow down further to say, okay, well, our analysis is only going to include those who gave us symptom surveys 12 months later, so that gets us to 2,600. So, for this very simple example, we might have 2,600 participants to analyze or samples and participants to analyze.

So, if we go to the next slide. Sorry, the next slide. So, changing this a little bit will of course change our numbers on the other slide. So, this is just switching it in a simple way. This is something we saw a lot in a row so far where say, okay, we don't want serum; we actually need plasma for our assay. Plasma was actually added a little bit later in the study. So, the numbers now for that initial sample go down to 2,200. And when you include then the surveys collected 12 months later, it's down to 1,800 for the analysis. This is showing a quick example of how using plasma is going to reduce your numbers. Each sample type we have is going to have little special cases that are going to change the numbers available. So, it's something to be aware of. As you're going in, you're just going to find the numbers will move around depending on what sample you need. And this is just one example of that. So, we can go to the next slide.

So going back to the serum assay, just as an example here. One thing we noticed as we work with our ROA investigators is there are of course many criteria you want to stack on top of that very simple initial example. So, I'm going to go through here some additional criteria just to show what the numbers might look like. So here we have the same initial question. You're running a serum assay on a sample collected within 30 days of infection comparing that to symptoms 12 months later. And then we have to exclude all participants with prior cardiovascular comorbidities. We have a comorbidities form that asks about this. So then after making that exclusion, the sample number goes to 2,000 and the 12-month surveys to use for follow-up go to 1,500. So, our analysis count is down to 1,500. If we go to the next slide.

And again, adding another criteria here. So, on top of the exclusion for cardiovascular comorbidities, we might say, and we only want to analyze people who had shortness of breath in that acute 30-day period. This will bring our sample count all the way down to 300. And then if we're including that 12-month required surveys, we get down to 222. So, this is what we see across the board. Every analysis we work with. There are lots of little criteria that are very important for any analysis. As they stack on top, you should expect, obviously, the numbers to reduce quite a bit. So, I just wanted to show an example of how our top line 15,000 participants with 140,000 surveys reduces down as you talk about your very specific analysis question. If we go onto the next slide.

Just want to pause to say a few general caveats that, again, that have come up with the ROAs quite a bit. As a reminder of what the study looks like as time passes, you're looking here comparing 12-month surveys to an initial infection. Well, they may have been infected in that 12-month period. How do you treat those participants? This is something you're going to have to keep in mind. We have participants at baseline who are close to an infection, who are confirmed by PCR (polymerase chain reaction) [test], but some also who are self-report only. So, you might want to consider how they were diagnosed with COVID initially. And again, this is a quite simple—compare 12-month surveys to month zero sample collection.

You may also want to consider the 3-, 6-, 9-, or 15-month surveys. So, we do have that longitudinal data. You may want to consider persistence. This will obviously lead to more complicated analyses. Expanding that 12-month window for surveys may also expand the number of surveys available to you. Maybe some people missed their 12 month but got a 9 or a 15 month. So being very conscious of how that longitudinal data is structured as well will be helpful. So, we can go to the next slide.

Okay. So, some other criteria that might move around those numbers significantly. The examples we're talking about so far required that sample collected within 30 days of initial infection. If we weren't interested in that sample period and we said, actually, we just want a sample from 12 months after infection, and we'll look at the surveys from that same time point. Well, we have much more participants here with that sample available, at 6,700 here, because we have lots of people who enrolled in the study months after that initial infection. So, the numbers are much higher here. We can move on to the next slide.

And then another thing that's going to change the numbers quite a bit is that what we've been talking about so far is biospecimen analyses. Maybe you do not have any need for a sample, so maybe you just want to look at surveys collected within 30 days of the initial infection and surveys 12 months later. And you're going to have a much larger sample size here because we have more surveys. So that initial number within 30 days is going to be 6,000. Sorry, there's a bit of a typo here. That should be participants with baseline symptoms form within 30 days of index, 6,000, and that only reduces down to 4,700 when you're looking at the 12-month visit. We can go to the next slide.

Okay. So just these are things to keep in mind as you're planning out your analysis question. Adding in more criteria will obviously reduce numbers, and this is to give an idea what that might look like. There are many, many other complexities in our study. I'm going to mention a couple extra ones here that will move those numbers around that are available for any analysis. If you're looking to analyze any Tier 2 or 3 testing data, these were rolled out over the last 4 years, so they weren't all initially available.

Participants are, as Leora mentioned, offered based on specific criteria. And then when the participant does a test, they're not offered again for a year. So that longitudinal nature of our data is a little different for the Tier 2, Tier 3 testing. Similarly, with lab results, you're going to see much less longitudinal data. And anytime your analysis requires lab results from our data, you're going to see numbers drop off quite a bit. But of course, if that's important to your analysis, it's just something to keep in mind as you're working through.

I think that is my last slide. Thank you.

Dr. Christine Bevc:

All right. Thank you, James. Thank you, Trisha. For our audience, James and Trisha are sitting right next to each other, and so that's why when Trisha was talking, you could see James because they're using James' microphone on that. But we want to make sure you guys don't get that feedback there. All right. So, we're going to turn over to the pediatric observational cohort study and Dr. Rachel Gross is going to help give us the overview there, and then we'll bring James and Trisha back in to give us a little bit more on the data side of these. So, thank you all for your questions. Go ahead, Dr. Gross.

Dr. Rachel Gross:

Oh, great. Wonderful. Thank you all so much. I am a general pediatrician and physician scientist, and I lead the pediatric observational cohort study. And we're really excited to share that RECOVER truly is a study that aims to characterize Long COVID across the entire lifespan and really does this in a comprehensive way for children across their pediatric lifespan. And so, we have similar study questions that RECOVER is trying to address to understand how many children are getting Long COVID, why do some children get Long COVID, and others do not. What symptoms do children feel when they get Long COVID and how long do they feel sick? What causes it to happen? As well as trying to understand how does having Long COVID affect later physical health, mental health, and development in children. Next slide.

And so, our pediatric observational cohort study has a different study design than the adult cohort. And we describe ourselves as a meta-cohort, really bringing together 4 different cohort types into one comprehensive study. And so, this includes first what we call the main cohort, which are participants that are newly recruited into RECOVER, and they span from birth all the way through young adulthood and include children and young adults who had an infection in the acute period, that had an infection a while before enrolling, as well as those that were uninfected. And we're also truly a dyadic study, and so we enroll primary caregivers as well.

In addition to this cohort, we have other existing cohorts that have combined and contribute data to RECOVER. One includes the Adolescent Brain Cognitive Development Study, or ABCD, which is a cohort of adolescents. Another is the MUSIC study, which is a cohort of children who had MIS-C (multisystem inflammatory syndrome in children) and are being followed longitudinally over time as part of RECOVER. And another cohort, which we call the in utero cohort, which are children born to the pregnant individuals in the adult cohort that may or may not have been exposed to SARS-CoV-2 during the pregnancy. And I'll give more details about these different groups. Next slide.

So, in general, we have children and caregivers who've enrolled those that have and have not had a history of a COVID infection, and we have those that have and have not developed Long COVID. And this spans the entire pediatric age group, from having infants and toddlers, preschool-aged children, school-aged children, teenagers, as well as young adults up to 25 years old. And we are a very diverse cohort, as you'll see in the subsequent slides.

So, we have over 22,000 participants who have contributed data to our pediatric observational cohort study across over 100 sites around the United States, including Puerto Rico. And we have about 8% that identify as Asian, 14% that identify as Black, 28% that identify as Hispanic, and 56% that identify as White. And so, we are a large diverse cohort reflective of the population here in the United States. Next slide.

This slide really summarizes our overall structure and aims for the pediatric observational study. So, we generally have groups—those that had a history of a COVID infection and those that did not have a history of a COVID infection. And we've spent substantial time characterizing Long COVID based on symptomatology by different age groups. And while we see that symptoms can be similar across ages, we see distinguishable differences for the child age groups that I mentioned, and variables related to that are present in the database. Building on this, we collect a variety of different kinds of data from biologic data to clinical data to social data longitudinally over time. To help us characterize different types of pediatric Long COVID, to evaluate risk and resiliency factors, and to identify the underlying mechanisms for these conditions, as well as to lay the foundation for the development of treatments over time. Next slide.

So, I'm going to give an overview of our study design, which does vary from what was presented about the adult study. So overall, our Tier 1 reflects a baseline assessment, which includes participants in the main cohort, ABCD, as well as MUSIC. A subset of them includes those that had their infection within 30 days of enrollment. Then a large subset of about 3,000 participants is chosen and selected to be followed longitudinally over time in order to oversample for children who have a history of Long COVID, as well as various comparative groups. So, they come in for in-person visits at 6 months, 12 months, 24 months, 36 months, as well as 48 months after enrollment. In addition, our Tier 3 includes more intensive studies for those most severely affected that occur 2 times in a subgroup of about 600 participants throughout the study period. Next slide.

So, as I mentioned, our ABCD cohort of adolescents participates in our baseline assessment because they are already a longitudinal study on their own, and they collect survey elements as well as biospecimens at baseline. Next slide.

Our MUSIC cohort, focused on children with a history of MIS-C, do modified versions of our different tiers to follow these children for longer, to give an opportunity to study the long-term impacts of having had MIS-C.

So, this gives an overview of our tiers. Our Tier 1 is our baseline assessment consisting of surveys and biospecimen collection. Tier 2 is our longitudinal study of a large subset of children who participate in both remote survey collection as well as in-person assessments, and our Tier 3 are in-person more intensive assessments, and I'll give a little bit more details about these going forward.

So all of our participants do our baseline Tier 1, which are remotely delivered surveys that include a wide array of information such as demographics, a child's global health, their past medical history, a lot of information about the COVID infection status, their vaccination status, comprehensive surveys about prolonged symptoms related to COVID and Long COVID, as well as assessment of child health more globally. So, impacts on diet, physical activity, sleep, school, and other aspects of child wellbeing, as well as social determinants of health, including blood and saliva biospecimen collections in both the child and the caregiver. Next slide.

Our Tier 2 longitudinal assessments and Tier 3 assessments are summarized here. So for our in-person visits, we also have a clinical exam that includes weight and height and other vital signs. We assess heart and lung function using ECGs and spirometry longitudinally over time. We have neurocognitive and emotional development assessments, as well as collecting biospecimens at each of these time points, both for biostorage as well as some tests being run at those times.

Our Tier 3, which are the more intensive tests in the subgroup, include assessments of cardiopulmonary function using echocardiograms, cardiac MRIs, cardiopulmonary exercise testing, pulmonary function tests, sputum induction tests, as well as neurocognitive assessments that include brain MRI, EEG, more comprehensive neurocognitive and emotional developmental assessments and additional biospecimens of other non-blood-related specimens.

This just gives an overview of the neurocognitive and emotional developmental assessments that we include in pediatrics because this is one of the primary aims of RECOVER as children grow over time. And all of these assessments are designed to be age appropriate across this pediatric lifespan. And so, for Tier 2, neurocognitive assessments primarily rely on PROMIS (Patient-Reported Outcomes Measurement Information System) measures as well as NIH toolbox measures to assess different aspects of development, including attention, executive functioning, memory, language, processing speed, and recall, as well as assessments of emotional development, including symptoms of anxiety, depression, stress, child behavior, and others.

And our Tier 3 neurocognitive assessments allow us to get more in depth about these neurocognitive and emotional development over time, looking at verbal and nonverbal skills, attention, memory, as well as children get older, assessing more academic skills such as reading, spelling, and mathematics, and more in-depth psychological symptoms and disorders.

And so, I'm going to end my talk just highlighting some of the unique differences about our final cohort type, which is the in utero exposed infant cohort. And so, one of the unique aims for this cohort is trying to aim to characterize the clinical manifestations of exposure to SARS-CoV-2 infection during pregnancy on child physical health and child development. And we do this by comparing children with and without a maternal history of SARS-CoV-2 infection during pregnancy. Next slide.

And this figure gives you an overview. We have pregnant individuals that are part of the adult RECOVER cohort who either had a COVID infection or did not have a COVID infection during their pregnancy. And we're following those babies longitudinally over time. Next slide.

And so, for this cohort, our tiers are described slightly differently, and that Tier 1 occurs longitudinally over time but includes remote data collection. And our Tier 2 recurs longitudinally over time but focuses on in-person assessments. And here we begin following these infants at 12 months of life, and we follow them when they're 12 months, 18 months, 24 months, 36 months, and 48 months old. Next slide.

And so similar assessments are done in this group, assessing for medical problems, a history of COVID infections and symptoms that develop. And social determinants of health. Next slide.

And what's unique about this cohort is there is more extensive developmental assessments that occur. So, this summarizes the ones that are collected remotely, including ages and stages questionnaires, assessments using the M-CHAT (Modified Checklist for Autism in Toddlers) screening tool for autism. We have an ages and stages social emotional questionnaire. As well as the use of the child behavior checklist to look at different behavioral problems. And the developmental profile for an interview assessing physical development, adaptive behavior, cognitive and social-emotional, and communication in children. And then our in-person visits include measurements as well as comprehensive developmental in-person assessments using the Bayley Scale of Infant Development and the Differential Ability Scale too. And we have one biospecimen collection in this group that occurs at age 24 months.

So, I will hand it back over to talk about more specific data structure for pediatrics.

Dr. Christine Bevc:

All right. Thanks, Rachel. And back to James with these for the data structure, which is a little bit different, as you might have figured. But we're going to run through these pretty quickly and try and stay on time today. So, lots of great questions. Thanks. Back to you.

James Chan:

Thank you, Rachel. Yes, I will go through these quickly, and I actually have much less here. I did not run through a full example, what we talked about before about the sample sizes and adding criteria, how it might affect the numbers you have available, obviously applies here as well. And we've seen very similar kind of things come up with the previous ROAs who have been working with the pediatric data.

Also, just as a response to some questions that have been going through the Q&A, I just want to emphasize that the NBR materials do also apply to pediatrics. And if you go through that site and get into that resource and go through the course, the data dictionaries and many of the materials are available there as well.

Okay. So, I'm showing the symptoms forms for pediatrics here in a very similar way to how we looked at it for adult. And then you can see, obviously, the counts here look quite different. Everybody baselined for us. So, we have 13 to 14,000 symptoms forms available at that time zero. Time zero here is different to adult. In adult, we always refer to the initial infection date as time zero. And for uninfected participants, that kind of negative test date that's stand in. In peds, we generally just refer to time zero as their first visit. People do come in at their first visit very close to an infection or a long time away from an infection. We'll look at that in the next slide, but I'll stay here for now. If they come in very close to their infection, they do a 2-, a 4-, and an 8-week check-in. The 8-week is included here, just to note that there is a sample collected at 8 week.

So, for the baseline participants, where baseline includes this 8-week for those, what we call acute participants, there is mostly symptoms forms and there is some sample collection. Technically, there is a general baseline sample collection, which is a dry blood spot, but I don't think it's going to be used for any future analyses. Okay.

So then after baseline, a select group of people are promoted into what we call Tier 2. And those people do an extra set of surveys. And overlaid on this, so everyone comes in and the surveys overlaid as the amount of people who end up doing the sample collection as well at each of those visits. Rachel went through what's done in each of these visits but just wanted to show what the counts look like here. And there is a big drop from baseline into Tier 2. We do expect month 24 and month 36 there to pick up over the next year as we continue the longitudinal data collection for pediatrics.

And I have not put Tier 3 into this. Tier 3 is an extra selection again of the Tier 2 participants. I'm not sure. They kind of happen concurrently with some of the Tier 2 visits and can also happen in between. So, I didn't want to include it on this kind of timeline thing. But there are 2 visits included in Tier 3 also, and it is again a reduced number of participants who are going to do those visits. Okay, we can move on to the next slide.

This is to show that one baseline bar that was in the previous slide for the post-acute participants, how far are they from a first infection when they come in for that baseline? And as you would expect, it's distributed quite well over time here. So many people come for the first time close to an initial infection but also several years out. We can go to the next slide.

This is the same data as in the first slide, just leaving at the baseline and showing the Tier 2. About 15% of the people brought into Tier 2 have no history of COVID infection at that 6-month visit. That of course may change as they move through. And I wrote a little bit here what's collected at these visits, but I think Rachel went over that much more thoroughly. So, we can go to the next slide, and this is the last.

So just some additional comments. I think I mentioned this a little bit already about Tier 3. Survey data, I'll mention this explicitly just because it was brought up in the comments. The survey data in adult and in peds are done by an online collection system through REDCap, so that participants can log in and answer the questions directly. In the case of participants in the pediatric study who are under 18, it's the caregivers who go in and answer those questions. And over 18, the participants themselves can answer the questions.

As Rachel mentioned, there is a separate in utero cohort. I didn't describe much about that here. Rachel thankfully gave some overview. Again, to get more into that, there are many materials available in NBR. And then I wanted to mention briefly that there are 7,000 participants who contributed a baseline from an external cohort that predates COVID. This might be interesting if you have specific research questions that want a more representative sample of participants.

And that's it for me. Thank you very much.

Dr. Christine Bevc:

All right. Thank you all for breaking that down for us. Let's bring in Emily and Cera to help us better understand where all of this data lives and how you can begin to access it. All right. Emily?

Emily Hughes:

Great. Thank you.

Hello, everyone. My name is Emily Hughes, and I'm joined by my colleague, Cera Fisher, today to tell you about NHLBI BioData Catalyst and how you can explore RECOVER datasets. Next slide, please.

All right, so before diving into specifics about accessing and analyzing RECOVER data, I first want to take a moment to introduce BioData Catalyst. Next slide.

BioData Catalyst is a cloud-based ecosystem of tools and resources to support research. The National Heart, Lung, and Blood Institute initiated this effort to address the needs of the research community for access to data, advanced cyber infrastructure, and leading-edge tools. There are various platforms in the ecosystem that support search, analysis, tools, and workflows to democratize data and computational access. All aimed at ultimately advancing science and to improve lives of patients. Next slide.

Today, we will be covering 2 main parts of BioData Catalyst or BDC. The first is BioData Catalyst powered by PIC-SURE, which I will be telling you about today. This is a search tool that allows you to conduct feasibility assessments of research hypotheses. This means that researchers can come to the platform with their questions in mind and determine which datasets would be helpful to find the answers to their questions. PIC-SURE also supports general data discovery and exploration, building cohorts of study participants, and preparing data in an analysis-ready format. The second part we'll discuss today will be BioData Catalyst powered by Seven Bridges, which my colleague Cera will cover later. Next slide.

So, let's talk about RECOVER data and how it can be explored on BDC powered by PIC-SURE. As I mentioned, PIC-SURE allows for data exploration and cohort building, but what does that actually mean? Essentially, researchers can search for key terms of interest across all BDC data, not just the RECOVER dataset. So, for example, I can search for specific symptoms related to my research idea. Like cough or head pain, or for demographics such as age. In BDC, we refer to these participant characteristics or attributes as variables. So that's what I mean in the future when I refer to variables in BDC. Once I find the symptoms, clinical outcomes, or terms of interest, I can apply filters on these variables. So as an example, I can filter to participants that had head pain or participants over 30 years of age. After applying these filters, I can get counts of participants that meet my filters or my query criteria. And if I have access, I can get the participant-level data to do my research. Next slide.

Our team has created a publicly available version of this tool that anyone can use. This is because data discovery can be difficult. Complex data can be hard to understand, and researchers often have a research question in mind but aren't sure how to find data that will support their idea or support their research for their question. BDC Powered by PIC-SURE is a publicly available cohort-building tool that is available to anyone. There's no login or registration required. This promotes equity to data exploration and allows anyone to explore data available on BDC, including the RECOVER datasets. So today I'll be demonstrating some of what is possible with this tool, which you can do yourself. Next slide.

It is important to understand how the data is organized, and this is a starting point that I urge all researchers to do when you're starting out a project and you have a dataset in mind. Knowing how the data is organized will help you to understand what is available and plan your research and plan your analysis that you have. So, let's talk about the adult and the pediatric cohorts in BDC PIC-SURE specifically.

For the adult cohort, the majority of variables or the participant characteristics are separated by months post-index and infection status. We can see this in the examples below that I've included on this slide, where problems thinking or concentrating are available for infected versus non-infected participants at 0 or 24 months post-index, just as a few examples here. For the pediatric cohort, there are 3 main sub-cohorts within the larger study. So, there's pediatrics main, congenital, and caregiver. These sub-cohorts can be explored using the consortium-curated facets in PIC-SURE, which I'll demonstrate in a moment here. Similar to the adult cohort, these variables are separated over time. However, pediatrics uses visit type, which you can see in the examples on this slide where there's a dry cough variable for the 12-month visit and the 24-month visit. Next slide.

All right. So now let's take a look at a real use case of a RECOVER research question on Discover, which is the public version of PIC-SURE that I'll be showing today.

For this demo, I'm going to say that I'm a researcher looking to see if the RECOVER adult dataset has data available and participants that meet my research idea. So, let's say I'm particularly interested in brain fog of those that were infected. So specifically, I want to see how many study participants were part of the infected cohort, how many reported brain fog at 3 or 6 months post-index, and had a brain MRI performed at 3 or 6 months post-index.

And so, on the next slide, I've included a video recording of this demo of the publicly available version of PIC-SURE and how I could use this tool to explore my research question. And so, if you'd like to follow along, I've included the link in the chat here. But yeah, go ahead and play the demo here.

Here, we can see the page when I first navigate to BioData Catalyst powered by PIC-SURE. We can see that I'm not logged in, but there are still some options I can take to explore data. Let's use the Discover page to conduct the feasibility assessment.

Here at the top, we can see the search bar. This can be used to search for clinical outcomes, phenotypes, or other terms of interest, which we refer to as variables. On the left-hand panel, you can see some faceted search options that can help narrow down the search results. Let's try to search for brain fog.

Behind the scenes in PIC-SURE, this is searching across all datasets in BDC for search results related to brain fog. I want to note that this is searching across all parts of the variable for relevant results, the variable name, variable description, values, and even study-level metadata. Here, we can see results that look promising. Some are from the adult cohort, some from autopsy, and some from pediatric. Let's narrow down the search results using the faceted search, clicking on the RECOVER_Adult option under dataset.

Now, we can see that the results were updated to only show search results from RECOVER_Adult. These results are looking promising. We can see with the variable name column that there are variables called problems thinking or concentrating or brain fog. The variable name column also shows us whether the variable is referring to infected or non-infected participants and at different months post-index. If we want to learn more about the variable, we can click on the row or the "i" icon in the actions column.

Let’s apply a filter by clicking on the filter icon in the actions column. Here for my feasibility assessment, I am interested in those in the infected cohort at 3 or 6 months post-index. I’m interested in those that have brain fog, so I will select the “yes” values. So “yes, and I still have it,” or “yes, but not in the last 30 days.” To add the filter, I can click the plus button. Here, we can see there are about 2,310 participants who are infected and had brain fog at 3 months post-index. If I want to learn more about the breakdown of this cohort, I can click to the variable distributions tool to get a better understanding.

This graph shows the variable distribution of the filter that I have added to my cohort. And here we can see that many more participants answered yes, and I still have it than those that answered yes, but not in the last 30 days.

As I mentioned, I am interested in brain fog at 3 and 6 months post-index, so let’s apply the filter for the 6 months post-index variable as well.

As we can see, the total count of participants went down to about 1,470 participants. This is because filters by default are added together, so this means that there were about 1,400 participants that had brain fog at 3 months post-index and 6 months post-index. However, we can change this by turning on the advanced filtering feature using this toggle. This feature is new and currently under development, so if you have feedback for the feature, please let us know using the link here.

We can change this “and” to an “or” to get those that had it at some point over time. Now we can see the number went up to about 3,451 participants that had brain fog at 3 months post-index or 6 months post-index.

For the next part of our research question, let’s try to search for brain MRI.

We can see many results returned here, including “was the brain MRI performed?” or “date of brain MRI performed.” To further narrow down these search results, I can use the consortium curated facets, which were created in collaboration with the RECOVER DRC. One of these facets is “test performed,” which returns variables about whether or not a test was performed.

Now we can easily find the variables for “was the brain MRI performed?”. Let’s add these variables for the 3 and 6 months post-index.

As we can see, by default, these variables were “and”-ed to the growing query, resulting in a filtered participant count of less than 10. But let’s change this to an “or.”

Now we can see that there are 81 participants or so that had brain fog at 3 or 6 months post-index and had a brain MRI performed at either 3 or 6 months post-index.

I hope this shows you the power of PIC-SURE in performing feasibility assessments to determine what data to use for your research.

Great. So now, as far as next steps go, you can try it yourself. I know that was a quick demo, so I’ve provided the link here and the link you can find in the chat. Again, this tool is publicly available for everyone. There’s no login or registration needed. So, you can go to BioData Catalyst powered by PIC-SURE, perform searches of interest, add filters, and explore the RECOVER data.

Now, I know this was a quick demo, so if while you are exploring, you realize that you need some assistance on using the platform, you can reach out to us using the contact link on this slide, and I’ll provide that in the chat in a moment as well. When you submit a question with this, it will submit a help desk ticket. And I want to mention that I personally respond to the questions about PIC-SURE, so this is a really great way to contact me directly if you need more support. However, you can submit general questions to this link, and our support team will direct your question to the right person to get you an answer. So, thank you so much for your time and attention today. I'll now hand it over to my colleague, Cera.

Dr. Cera Fisher:

Thank you so much. I really enjoyed that presentation on PIC-SURE. My name is Cera Fisher and I'm the program manager for BioData Catalyst Powered by Seven Bridges. And I'm going to spend the next few minutes giving a quick overview of this analytical platform and describe how researchers can use it for deeper analysis of the RECOVER datasets. So next slide, please.

So, I'm talking now about the second BioData Catalyst platform that we are discussing today, which is Biodata Catalyst Powered by Seven Bridges. We usually shorten this to BDC-Seven Bridges. I want to apologize in advance because I will be using a little bit more technical research jargon than Emily did in her discussion. And that's because Seven Bridges is a cloud computing analytical space for researchers. Seven Bridges does require registration and access to the data does require an approved registered research plan. This is because on BDC-Seven Bridges, you can access and analyze the participant-level data, which requires secure workspaces to protect individuals' privacy.

On BDC-Seven Bridges, researchers can set up secure workspaces that allow them to collaborate with researchers across different institutions while staying inside a regulatory compliant environment. Researchers are able to analyze the data where it lives rather than downloading it locally. You can set up analyses using either a graphical user interface, or for researchers who are more code savvy, you can use an API. Interactive data analysis can be done using JupyterLab notebooks for Python programming, RStudio and SAS Studio, all of which provide high-level statistical analysis options. The platform allows researchers to access top-of-the-line computer hardware through AWS, that's Amazon cloud computing, or through Google cloud computing. There are associated costs for using cloud computing, which are directly passed through, but I will talk a little bit about support to defray some of those costs at the end. Next slide, please.

For researchers, here's what the computing environment gives you. As I mentioned, you analyze the data where it lives. The data stays within a secure environment protecting privacy. You do not have to worry about making sure that your own local computing environment meets all the regulatory requirements for working with sensitive personal data. We have done that work. BDC-Seven Bridges democratizes access to high-powered computing, and this is particularly important for researchers who are at small colleges and universities. At big R1 research universities, researchers may have access to a university-owned data center, but smaller schools don't necessarily have that. However, with BDC-Seven Bridges, if you're a researcher with an approved research plan, all you need is an internet connection. You can get started analyzing the RECOVER data.

There's no computer system that the user has to manage. You don't have to install software, download files, and get everything in place. You can just log in and start analyzing. You also only pay for what you use. The most common resource is 34 cents per hour because you're renting time on advanced hardware rather than having to purchase the whole computer. And we have 24/7 technical support, scientific research support, and robust documentation to help you with your work. Next slide.

I'm going to give an overview of the 2 ways that registered researchers can access RECOVER data in BDC-Seven Bridges. You can access the RECOVER data through a tool called Data Browser in the visual graphical user interface, or you can start at PIC-SURE, as Emily was showing, use those tools to search and filter to build your subset of the data that you intend to analyze, often called a cohort. And then, you can use tools to import the PIC-SURE cohort into Seven Bridges for analysis. Next slide.

Please do not think that this presentation is meant as the only tutorial you can get. This is just to give you some insight into how a researcher can access the data. When you're ready to get started, we will be here to help you. To use the data browser, you will log into BDC-Seven Bridges, you'll click on "Data" and then click "Data Browser." You'll see a list of all of the datasets that are hosted in BioData Catalyst and be able to select your RECOVER datasets of interest. As a side note, one of the many benefits of using BioData Catalyst for researchers is that you can co-analyze the RECOVER data alongside, for example, other COVID datasets that are part of the NHLBI-hosted data like C4R, the Collaborative Cohort of Cohorts for COVID-19. After you've selected the datasets you want to work with, you can copy files of interest to your project workspace, and from there, you'll begin analyzing the data. So that's one way to get to the data. Next slide, please.

The other way to get to the data is to import cohorts from PIC-SURE. If you've created a cohort with your filtered study variables at PIC-SURE, you can import that data to BDC-Seven Bridges with a provided script. You would navigate to "Public Resources" and then find "Projects." You'd open the project data export from the PIC-SURE UI and then navigate to the "Data Studio" tab. You'd be able to copy the data studio analysis to your own project workspace and then follow along with a beautiful script that Emily wrote for us that will teach you how to import the data. Next slide, please.

For many researchers, moving from analyzing data locally on your own computer to analyzing it on the cloud is a bit of an adjustment. We are here to help. As I mentioned, there are costs associated with using cloud computing, and the NHLBI offers new users $500 in pilot fund credits to get started. Five hundred dollars actually goes a pretty long way. I know some researchers who have completed their projects within that budget, but most importantly, this means you're not going to have to use your grant dollars just to learn how to use this system. BDC also has a YouTube channel with demos and instructional videos. I have a link that I will put in the chat of Emily demonstrating how to import data from PIC-SURE to Seven Bridges. We also have troubleshooting and technical support available 24/7. You can contact our support team by sending an email to support@sevenbridges.com.

For live support, Seven Bridges has drop-in style office hours twice a week. You can chat with PhD scientists and get not only technical support, but also research support in designing workflows, accessing data, and using the tools. And we have a link here in the slides for you to learn more about that. Please don't hesitate to reach out for more information, and thank you so much for being here today. That's all from me.

Dr. Christine Bevc:

Thank you, Emily. Thank you, Cera, and thank you to our presenters for just sharing all this information. Now, we've got some time to move into the Q&A portion of today's seminar. We're running a little bit tighter than we'd like. So, if we don't get to your question, responses will be provided along with a summary of the seminar on the website. Anytime you can also look over at the Q&A to click on the "answered" tab to see if your question has been answered with a written response. And there's 25 answered questions over there. So, everyone has been really busy and we appreciate that. But I want to start us off with a question we've seen a lot in the chat there and received many times with the pre-submitted questions, which is, "How do I access the data and repository?" And this is a great question.

And yesterday, all of the registrants, you should have received 2 PDF files. The file titled Data Access and Availability Guide is going to be your starting point to learn how to access the RECOVER data and which types of data are going to be available for study through the NHLBI BioData Catalyst. So, the link also for PIC-SURE is in the chat there. Thanks to Emily for sharing that. And you can get started with that today to be able to start drilling down on that. For the approved research plan that Cera mentioned, for access to the BDC Powered by Seven Bridges, if you submit an application to the research opportunity announcement that I mentioned at the top there, and if you're selected for funding, your research plan is approved through that process, and you'll receive instructions for accessing. If you have a research plan, you could also submit your request to the RECOVER's ancillary studies. So that second document that you received, Research Opportunities and Data Access Flowchart, is going to help guide you through that process there to figure out how to submit your research plan to be able to request access.

So please look at both of those documents and submit any other questions that you might have through any of the resources that our amazing panelists have provided, and they're just continuing to drop in those links into the chat there. So, all of these will be available. Our recording is posted to the RECOVER website after today. So, if you miss something, don't worry, it's going to be available to you. Now, we have another question, which we've seen a lot also over in the Q&A. And that question is, "How can I find out exactly what data is available for a specific test? Like how do I find out if olfaction has been tested? Was it tested longitudinally? Where can I find mental health?" So, Emily, I want to start with you because it feels like PIC-SURE is going to be a good place to get started. We saw the demo; can you just remind us how folks can get started?

Emily Hughes:

Yeah, absolutely. So, I dropped the link in the chat to BioData Catalyst Powered by PIC-SURE. And if you go over to the Discover page or the Discover tab, or you can even use the search bar that's right on the homepage there, you can search for different terms of interest. So, if olfaction tests were something you wanted to search, you could search that. And maybe a little pro tip, I guess, for searching for data in PIC-SURE, I really encourage you to try a couple of different search terms to see what's returned across those terms. So, someone was asking about loss of smell, which might be related to what you're trying to find with olfaction tests. So, try a couple of different searches to see what comes back and what data is available.

Dr. Christine Bevc:

And specifics, so this is a question actually over to Leora and Rachel around what exactly are the specific tests that have been conducted? Where can they find the operations and details on the protocol?

Dr. Leora Horwitz:

Oh, well, the protocols are all posted publicly on the RECOVER website. I'm sure someone can drop the link in the chat here. So, you'll always be able to see exactly what we've been collecting in the study.

Dr. Rachel Gross:

I also wanted to highlight that each of the observational cohorts also has a study protocol manuscript that was published that might give an overview of the study design and all of the assessments that were included. And then, the most up to date would be the protocols that are shared online that Leora mentioned.

Dr. Christine Bevc:

Great. There's lots of resources available on there. And then, as James and Trisha mentioned, there's also the NBR, the Network for Biostatisticians for RECOVER, and those courses there. Thank you, Brian, for just dropping that link there into the chat. Much appreciated there. One of the next questions that was around BDC, so this is going to go to you, Cera. "For investigators, oftentimes they're collaborating, they're working with others. Is there a way that they can set up a shared space in BDC to be able to work across? Because you mentioned the data privacy, the data security, so what if there are different institutions?"

Dr. Cera Fisher:

Thank you so much for this question. It is one of, I think, the strengths of BioData Catalyst are you are able to set up secure workspaces that you can invite other researchers into, and you can set role-based permissions that give them different levels of access to copying files or executing workflows. That's all handled through a registration process. You would need to have an eRA Commons login to be able to use this service, but it does mean that researchers can really easily collaborate with each other, share information, share data, share scripts without having to resort to something that you're not allowed to do, which is email each other Excel files. So, it's one of the things that I really like about the platform.

Dr. Christine Bevc:

That's great. And then, as a follow-up question, and this is one that was just submitted there, is for the computational cost: "And it's great the $500 getting you started, but how can investigators, potential researchers find out how much it's going to cost to run?"

Dr. Cera Fisher:

So, I'm not going to pull a punch here. It's not necessarily easy, but we know that. We know that you're going to need support to figure out how to assess your costs. All of the costs for using the computational services are the direct costs from Amazon Web Services or Google Cloud. Those costs vary depending on how large a computer you decide you want to run on. So, if you want to use more RAM, if you want to use more CPUs, it will cost a little bit more. Generally speaking, for working with the RECOVER data, you're probably going to be needing to use a resource that costs about 50 cents an hour.

And so, you have to figure out, "How long is it going to take me to set things up and how long is the actual analysis going to take to run?" But those costs are going to be stable at that per hour charge. It's prorated by the minute. You can easily get into real weeds here and I'm going to stop talking so I don't take up everybody's time, but there's lots of support to help you figure out how much you're going to spend.

Dr. Christine Bevc:

Great. And I think we just had one of your colleagues submit some information into the chat there, so we'll funnel that back out because that does provide the details in terms of estimating and managing costs, which is great to hear. So, let's see, we've got time for about one more question, and I'm going to take an easy one here because for investigators wanting to know again about the data, where it is, and James, I think this probably goes to you and Trisha on this one, but is there a way to access the codebook? And I know the codebook's pretty intense, but what would you suggest for investigators that want to get in there and see that?

James Chan:

I would recommend the NBR website again. Lots of materials on there that'll include the data dictionaries or codebooks and lots of material and lots of resources around it to help you understand them as well.

Dr. Christine Bevc:

I'm going to squeeze one more question in here, and this is probably going to go back to you, Cera. "Is there a way for investigators to also import data into BDC to analyze the RECOVER data with another cohort that they have?"

Dr. Cera Fisher:

Yes, absolutely. Multiple different ways. Not going to go into it, but yes, absolutely. You can co-analyze the data with other data that you might have that's pertinent.

Dr. Christine Bevc:

That is a wonderful tease for possibly a future R3 seminar. Great. Well, thank you everyone for all of your questions. We have received tons of those into the chat there. And thank you, Brian, for just posting that link to the data dictionaries there. Today, any of the questions that we receive are going to be shared back out. We take the names off of those, so it's just the questions that you ask, not any of your identity associated with it. If you ask specific questions about cases or analyses, those are excluded because we're unable to answer those here, but you should look for that recording of today's seminar. It's going to be available on recovercovid.org in about a week or so. And then, we'll also be posting that summary along with the Q&A. So that's also going to include all of the responses to the questions that were submitted. So, with our Q&A, we've got over 30 answered questions in there that are going to be showing up there. So definitely check back about that for that.

We thank you so much for joining us today. We hope that you'll be back for Part 2 of this series next month. We're going to be focusing on the biospecimens and the samples collected, the frequency of those samples collected, where to find details about the collection process for the RECOVER observational studies, so adults, pediatric, and also we'll be covering autopsy as well. So actually, you should also see, lastly, the short survey that's popped up on your screen. We're asking for your feedback on today's seminar as well as future seminar topics that you'd like to see. So please take a moment to fill this out. It'll help our planning committee learn best how to plan moving forward. And we just overall want to thank you for joining us today and hope that you have a great rest of your day. And we will see you next year. Thank you.