Transcript

Christine Bevc:

Welcome, everyone, to the RECOVER Research Review, or R3, Seminar. My name's Christine Bevc. I'm a translational health scientist with the RECOVER Administrative Coordinating Center, and I'll be your moderator for today's session. The goal of this seminar series, for those of you joining us perhaps for the first time, is to catalyze a shared understanding of the research within the RECOVER Consortium.

I want to start by thanking everyone who submitted questions in advance, and remind everyone, as Cameron mentioned, that you can submit any questions today during the presentations using that Q&A feature in your Zoom menu. After today's panel, our speakers will be addressing as many questions as possible.

We'll also have a Q&A document that's going to be posted, along with the recording of the seminar, on the recovercovid.org website. The document will contain the answers to questions submitted, relevant to today's presentation. Questions about other scientific topics will be addressed in future seminars, and answers to broader questions about RECOVER will be available in the FAQs found on recovercovid.org as well. As a reminder, we cannot answer questions about individual clinical care. Next slide, Cameron.

I'm pleased to share that our presenters today are Mr. Chris Lunt, Dr. Emily Pfaff, and Dr. Hiral Master. I think we can go to the next slide, so we can see our wonderful panelists here.

Chris Lunt is the Chief Technology Officer for the National Institutes of Health, All of Us Research Program, where he leads the building of platforms that engage participants and engage researchers to accelerate precision medicine. His experience spans work in government and commercial sectors, and including founding venture-backed startups. He's a lifelong learner, and also hosts a science book club.

Chris is joined by Dr. Pfaff, who is an assistant professor in the Department of Medicine at the University of North Carolina, Chapel Hill, School of Medicine, and serves as co-director of the Informatics and Data Science Core at UNC's Clinical Translational Science Award Program. Her primary research interests are clinical data modeling, clinical data standards and harmonization methods, and computable phenotyping. Dr. Pfaff is also one of the leads of the National COVID Cohort Collaborative, N3C, which is one of the largest harmonized clinical datasets in the United States. And she also serves as one of the MPIs of the EHR, electronic health record, and real-world data component of the NIH RECOVER Initiative.

And rounding out our panel is Dr. Master, a licensed physical therapist and certified public health professional. She holds a PhD in clinical and translational research and serves as the research data innovation lead for the All of Us Data and Research Center. She leads efforts and phenotypic and digital health data to drive scientific innovation for the program. With over a decade of experience, Dr. Master has worked extensively in clinical research, overseeing multidisciplinary teams and managing complex health studies. Her passion for data-driven healthcare and promoting diversity in research drives her mission to improve healthcare delivery and population health outcomes.

The topic for today's seminar is Advancing Long COVID Research by Fostering Collaboration between RECOVER and All of Us. Today's speakers will each provide an overview of the National Institutes' All of Us program and some of the leading-edge science that has evolved from it, including the analysis to reproduce the N3C RECOVER long COVID phenotypes. The presentations will also share information about All of Us data availability for long COVID research. And with that, please welcome all of our speakers, and I'll turn things over to our first presenter, Chris.

Chris Lunt:

Thank you so much, Christine. Hello, everyone. I'm very excited to be here and talk today about the program. Next slide, please.

And about this collaboration, which I think was a great example of what we can do when we are working together across these very large programs. Next slide.

And in this case, it was RECOVER, All of Us, N3C, PCORnet, and then Deloitte was helping with some of the organization. And this was a really exciting group of scientists to get to work with, and very high-profile public health programs, demonstrating how you can work together to advance the science. Next slide, please.

And so, the goal of the project was to come together to help support more research efforts to develop a better understanding of long COVID. And it had started with Dr. Pfaff, with a team, had developed a ML model to identify patients with long COVID. We saw an opportunity there to use the All of Us Research Program dataset to be able to do real-world testing of that identification model and to bring an audience that was diverse, and to bring an independent dataset to that. Next slide, please.

And so, we kicked off this effort between these different groups, using tools that have been developed across multiple teams to test machine learning interoperability across multiple platforms that had been built, and have to deal with the policy and access issues that that would create, and to, in general, look for ways that we could combine our expertise to improve our understanding of the kinds of analytical approaches that you can use in these cases. And we kick this meeting off... The kickoff meeting, to start this all, was in May 2022. Next slide, please.

But my role here really is to introduce you to the All of Us Research Program, about what we've done and the data that we've collected, that we brought to the table to support this effort. Next slide.

So the All of Us Research Program really goes back to the first sequencing of the human genome. And at the time, the director of the National Human Genome Research Institute was Dr. Francis Collins, and he saw the opportunity that this capacity to sequence human genomes was going to bring to medicine and to really enabling more precision medicine. But he felt that the way that we were going to be able to understand and read that large book of As, Cs, Gs, and Ts that we were able to produce, and understand what that meant in terms of health, was to engage a really broad audience of people to talk about their experience [inaudible 00:06:52] that we could then correlate with what we were seeing in the genome. So everything that we've been able to do in All of Us has been thanks to the participants and the community partners and their contributions to the data that we've collected. And that was one of the pillars of the program was to say upfront that we were going to treat the people who worked in our program not as subjects, but as partners and participants. Next slide, please.

And so, the goal of the program, which was then funded originally in 2015, was to get at least 1 million people from across the United States, and to be able to bring in the full diversity of the US, to bring in communities who were underrepresented in biomedical research traditionally, and particularly, people who are from underrepresented racial and ethnic minority groups. Next slide, please.

And so, I talked about one of those pillars earlier, about really engaging with participants as partners. And so, we were also then really looking to engage communities who've been left out of medical research in the past, to be able to bring in a broad set of biological factors and social determinants on a large scale, to follow the participants longitudinally over time, and see how they move, age, and grow, and to make that easily accessible so that any researcher could come in, regardless of the funding or infrastructure that their institute had. Next slide, please.

So let me talk about where we are right now in that process. So since we've started, and we started doing enrollments in 2018, and we have released our first dataset in 2020, and we are now at 838,000 participants who are somewhere in the pipeline, people who've registered with us. And of those, more than 569,000 are participants who have completed all the initial steps of the program, which means that they have given us a bio sample, that they have filled out basic surveys about their demographics, about their health habits, about their family health history, that they have given us some physical measurements at the time that they enrolled, which included hip and waist circumference, height and weight, blood pressure, et cetera. Those bio samples, by the way, are blood ideally. In some cases where we can't collect blood, we do collect saliva, and then we're also collecting urine as well. And they've also given us permission to access their electronic health records. And at this point, we have more than 450,000 people who have offered their electronic health records to us.

As I'd mentioned before, we're looking to really try and engage a really diverse set of participants as well. And so, beyond the 50% or 45% that we aim to get, who are underrepresented by virtue of their race or ethnicity, we're also looking to bring in people who are diverse for many different reasons, including whether or not people who traditionally have had access to healthcare, people who are low income, people with disabilities, sexual and gender minorities, et cetera, including looking for people who are from geographies that have traditionally not been able to participate in research. And as you can see at this point, we do have participants from all states and from all territories as well. Next slide, please.

We have largely been hitting those goals in terms of people's bringing in underrepresented audiences in the curated data repository. So this is the data that is now prepared and available to researchers. At least 74% of the participants have at least one thing that marks them as being underrepresented, 43% have non-white race or Hispanic, Latino ethnicity, 24% are greater than age 65, 9% have less than a GED for high school education, 25% have an annual income, less than $25,000, and 10% are sexual and gender minorities. Next slide, please.

So we take all of this data from all of these different sources, which includes working with hospitals, like federally qualified health centers, academic medical centers, and the VA, from participants who've also offered to connect devices that they own into the program, like Fitbits. We harmonize all of that data. We do data refinements, and we do stripping of anything that could allow the participants to be re-identified. We put that into a curated data repository and make that available with a data dictionary. Next slide, please.

And so, the current version of the curated data repository that's available for researchers to use is CDR 7. We're on our seventh version, and this version has more than 413,000 participants who have survey responses, more than 287,000 people with electronic health records, more than 337,000 people with physical measurements, and more than 300,000 people with genotyping arrays. But perhaps more exciting is the fact that we have more than 245,000 people for whom we have done whole genome sequences, and that is with short read sequences, although we've also done a set of long read sequences as well. We have more than 1,000 people where we've done long read sequences, which helps find things like structural variations. And we have nearly 100,000 people where we have produced structural variant data. So this is things like inserts and deletes and transpositions and copy number variations, and things like that.

We also have more than 15,000 people, where we have made their Fitbit records available. And for the first time in this dataset, we also included sleep data. And that's a really exciting dataset, and one that Dr. Master has also worked on, since researchers have not generally had a lot of access to sleep data in the past, and certainly not at this scale. I'll also mention that we are on the cusp of releasing the next version of this dataset. And in that case, there's going to be some real improvements, and some of these numbers, we're going to, in fact, pretty much quadruple the amount of Fitbit data that we have available in the next dataset. Next slide, please.

And so, in terms of access, we have access at three different levels for everybody. First off, we have a public tier that anybody can come look at the data, and there's a data browser to be able to do this. In fact, I was just speaking to a researcher who's doing allergy research, and she was asking about how much data we had for people with peanut allergies, and I showed her how she could just go into the public tier and look for how many people had peanut allergy, either in the diagnosis as part of the electronic health record or also in labs and measures. And there were substantial number of them. She was very excited about that. And so, it's a great way to start if you're interested in the dataset to understand if there's a population that's large enough that you would likely be able to use them in a research project that you're doing.

This is all then those summary data. It doesn't include any row-level data in the public tier, and that's to protect the privacy of our participants. But we also do have a research projects directory, where you can see what projects other people are working on, a list of publications that we know about that have come out of our data, some data snapshots, and a survey explorer that allows you to look at what are the questions we ask in the survey explorer, and data snapshots to give you some highlights of what we have as data.

If you registered to use the Researcher Workbench, you'll have access to either the registered tier or the controlled tier, where you'll also have access to a cohort builder that allows you to build a list of people you're interested in, based on characteristics of the data that they have, to build the specific data that you want to use. And then you're able to pull that into a Jupyter Notebook, where you can use R or Python to work on that data. We also do have SAS available and RStudio if people would like to be able to work on the data that way. And we also have some workflow management for people who are doing genomic analyses.

And finally, we have the controlled tier, and what the controlled tier adds on top of the registered tier is the genomic data, most importantly, which is not available at the registered tier yet. And the only thing you'd need to do additionally, in terms of using the controlled tier, is go through a little bit more of responsible conduct research training that we provide, which is only 45 minutes to get done. And next slide, please.

At this point, we have more than 12,000 researchers who are registered on the workbench, from more than 890 institutions. This includes minority-serving institutions and historically Black colleges and universities. We've also made the data available for the first time internationally last fall, and we have just opened up to commercial access by commercial organizations as well. There are more than 12,000 active projects in there. We are more than 600 publications and peer-reviewed journals that have come out at this point, including 300 this year alone. So the pace is really accelerating. It's getting hard for me to keep up on trying to read everything that's coming out of the program. And we're seeing some common conditions that are being included in a lot of these studies, including cardiovascular disease, COVID, mental health, cancer, diabetes, and health disparities. Next slide, please.

Yeah, I mentioned that there are a lot of publications, and here's a graph to show you what that looks like. You see a real acceleration as people are really digging into this data, and we're excited to see this start to turn into the possibility of greater precision medicine for everybody. Next slide, please.

That's it for my slides. I'm happy to answer questions, either in the chat or at the end, but now I'm going to hand it over to Dr. Pfaff. Take it away.

Dr. Emily Pfaff:

Thank you so much, Chris. And thank you for having me here at the R3 Seminar. I am going to spend my time here talking about how we can use algorithms to identify patients with long COVID within the EHR, which is the first step in being able to use data in order to research long COVID. There's a lot of promise to this, but there's also a lot of challenges, and I'm going to touch on both. Next slide, please.

So just as a bit of background, the data that I am going to be talking about today comes from N3C, or the National COVID Cohort Collaborative, which some of you may be familiar with. N3C is the largest national, publicly available HIPAA-limited dataset. HIPAA-limited, meaning that we have dates available which are considered a HIPAA identifier, so dates of birth, dates of death, dates of healthcare encounters. And then, we also have zip code data available, so that geographic-based analysis can be completed in N3C.

We are a representative cohort, certainly, as you can see on the map, that we do not have every state, and I want to be really transparent about that. But our demographics look very much like a breakdown of the United States, and we're really proud of that and feel like we are able to do really representative research, if not fully representative research, using that data source.

The data are harmonized, meaning even though we get all of the data from many, many different institutions, and it all gets mixed together in one big dataset for people to use, we've taken pains to clean those data and transform them all to the same format, so that when researchers use the data, they can query across all those different healthcare institutions, more than 84 of them, as if they were one giant healthcare institution. And that really helps with national scale research. And, of course, COVID is a national issue, a global issue really.

We also have linked data in N3C. So we have, for example, mortality data linked into the N3C dataset, meaning that folks who pass away outside of the hospital. That's often something that can be very difficult to ascertain in EHR data because the EHR really mostly knows about things that happen inside the hospital. But, of course, mortality is a very important research endpoint in both COVID research and lots of other kinds of research. And so, the ability to link with national death records, which is one of the datasets that we bring in, really gives us a lot of opportunity to have a more accurate view of vital status. And because we have this nationally representative dataset, we are able to see patterns and trends in disease, in this case, COVID, in a way that's not possible in smaller datasets. And so, all of the work that I'm going to be talking about today uses this N3C data. Next slide, please.

So many of you are no doubt aware that using data to study long COVID both has a lot of promise, but also a lot of issues. And I want to spend my time here being really transparent about those issues because I think, after a few years of working on this, we have really started to see the disconnect between the patient experience with long COVID and then the data that ends up being recorded in the medical record, which doesn't always line up. So in my Venn diagram here, you can see that patients, they care about the symptoms that they're feeling, obviously. They certainly care about the quality of life or changes to their quality of life from long COVID. They may have gone on what is known as a diagnostic odyssey, where they're going to specialist after specialist, different doctors, until they get somebody who will validate their concerns and get them to the right diagnosis.

And then there's this idea of improvement. Are people getting better? Are they staying the same, or are they getting worse? All of those things and many, many more things that I didn't list here are critical to the patient experience. And very few of those things are recorded in the EHR in a satisfying way. On the other hand, there are things that are in the EHR that aren't directly connected to the patient experience but can be helpful. So we have things like diagnosis codes that get entered in a structured way in the EHR and allow us to study what illnesses and symptoms a physician or medical practice has entered in the data for patients. That is absolutely not a complete list of symptoms, but it does give us a good starting point to work with.

We can also see what healthcare utilization looks like, so we can see what services were used, if procedures were done, or things like tests and imaging. Those are all going to be reflected in billing data and can be really important to help us see what care looks like for long COVID. Lab results, as we continue to search for signals, biomarkers, anything that is on the mechanistic side about long COVID. Those lab results are very voluminous, as you can imagine. There are literally billions of rows of lab results in our data, and we spend a lot of time mining those to look for clues and patterns. And then, of course, things like medications and vital signs are also critical pieces of the puzzle that are available in the EHR.

So with that background, note that many elements of the WHO and the more recent National Academies long COVID clinical definitions are inconsistently available in the EHR. And we can talk about this a little bit more if there's interest during the panel discussion, but there's a lot of interest, understandably, in enabling us to try and actually implement the WHO and NASEM definitions in our data to see what we get. But if there are elements of those definitions that are very symptom-based... For example, abdominal pain, that's just one very small example, but that is something that may not be reflected in the EHR in the way that a chronic condition like diabetes would be. And so, having it not in the EHR doesn't mean that the patient is not experiencing abdominal pain, and it doesn't mean it's not important, but it does mean it's not there for us to work with. And so, trying to implement those definitions has been challenging, not impossible, but challenging.

The other thing to note is that particularly important features of long COVID, particularly those that really affect patient quality of life to an enormous degree, like brain fog and post-exertional malaise, do not have specific ICD-10 codes. Or if symptoms like this have gotten ICD-10 codes since the beginning of the pandemic due to the hard work of patient representatives and advocates, they've only been around for a portion of the pandemic and are missing for a large part of our time period. And then, I caveat that we always use with EHR data research is that the EHR data is only available for patients who have access to and seek care. So just as much as we need to study the patients who are in our data, we also have to remember what patients are not in our data.

However, we do believe that there are a lot of affordances to EHR data to bring something to the table for long COVID research. One, as I mentioned, is our large diverse population. So we have over 23 million patients in N3C from many places across the country. And then, this concept of longitudinal data. So we have data that spans from 2018, so well before the pandemic, all the way up through the present, and that gives us the opportunity to really investigate trajectories of patient health over time. Next slide.

So let's talk a little bit about the diagnosis code for long COVID, which is U09.9. For a while, with EHR research on long COVID, we were working without any diagnosis code whatsoever, and long COVID research in the EHR felt like this insurmountable ice wall where we couldn't get a hold on anything because we didn't have a piece of structured data to hold onto and help guide the rest of our efforts. U09.9, when that came out and was made available for use on October 1st, 2021, gave us something to grab onto. So there was a lot of initial excitement when that was released. As you can imagine, when a new diagnosis code is released, it's not as if everybody starts using it that day. There is a period of uptake. So there is sort of... If you look at a graph of how much that code is used, you'll see this rise over time, slow rise, as that code is taken up more and more and more.

However, it does have some disadvantages, and it is not a perfect way. It's far from a perfect way of truly identifying who has long COVID data, and I'll talk through what that looks like. However, it is very useful to still characterize patients who do have that code because we can see other patterns that co-occur with that code. For example, other diagnoses that occur alongside a U09.9, medications that people are taking can be useful in studies of potential preventatives or treatments. And again, imaging and other kinds of tests that are happening around patients with that code can give us a good idea of healthcare utilization. So it is certainly not a perfect system, but it did give us something to start with. Next slide, please.

But let's talk about the disadvantage. Let's say that my goal as a researcher is simply to use any patient that has a U09.9 code as the cohort of patients that has long COVID. The issues that I would run into are a equity and access issues. So colleague of mine and others from the N3C team found that the patient group that has U09.9 codes on their records is disproportionately white and non-Hispanic, and is significantly less likely to live in areas with higher social deprivation. This does not mean that we believe that those individuals who don't have the code don't have long COVID. It means, in our hypothesis, that they have not been given the opportunity to get the code on their record. And there's a big difference between those two things. So if we only used U09.9 to identify these patients, then we would be certainly leaving people out.

There are also usage issues. We have looked at how U09.9 is being used in medical records, and there are some strange patterns that suggest, particularly in the beginning of its availability, that it was actually being misused. You can imagine that there's some confusing guidance sometimes for how a new ICD-9 code, or ICD-10 code rather, should be used. And if that happens, then perhaps a clinician accidentally codes U09.9 for acute COVID instead of long COVID. Issues like that that can cause odd patterns in the data that don't actually look like long COVID to us. For example, when a long COVID code is issued on the same day as an acute COVID diagnosis, it's not necessarily incorrect, but it does raise questions.

And then maybe most importantly, there are timing and coverage issues. So there are plenty, many people, many patients who ended up with long COVID before October 1st, 2021. And so those patients just missed out on the opportunity, just by virtue of timing, to get that code on their record unless they continued care after that date. And so, particularly from that early part of the pandemic, we just want to make sure that we're not instantly assuming that no one from that time period has long COVID by relying too much on the U09.9. And I think that the thing that really speaks to this is that if we assumed that only patients with the U09.9 code had long COVID, the data in N3C would show a long COVID prevalence among patients who had COVID of 1.2%. And based on everything that we've seen, even though there isn't a firm number that is the consensus prevalence of long COVID, we do know that 1.2% is misleadingly low. And so we know that that's not sufficient. Next slide, please.

So where does this leave us? We know we can't just use U09.9. And so, what the folks at N3C decided to do was to try to tackle this using machine learning. The idea is that we find the patients who do have a U09.9, of which there are tens of thousands in N3C, and we gather up all of their data. We gather up their longitudinal conditions, their medications, their lab results, everything that we can find about them. And then, we have a machine learning model look at all of that data, which would be too much for a human to consume, and try to learn patterns, patterns of data points that seem to indicate that patient A may be more likely to have long COVID than patient B.

Once we've trained the model on those patients that have U09.9, we then take that trained model and apply it to all of the other patients in N3C. And what the model is looking for are not patients with a U09.9. They're looking for patients that match the patterns that were trained into the model during the training phase. And in that way, we're able to identify lots more people who are potential long COVID patients than we would have just using the code.

When we open the hood on our model and look at what the important features are or the important parts of the pattern are, that the model is really hanging onto when it makes its decision, it includes things like healthcare utilization, do you have a lot of doctor's visits, shortness of breath, your age, certainly fatigue and cough, any CFS, gender, heart palpitations, pain, lots of different respiratory issues, and sleep apnea. And there are probably about another 180 or so features beyond the ones I have listed here that have some importance to the model's decision-making. But you can see that that list of features is pretty expected. There really aren't features there that make us tilt our heads and wonder where they came from. A lot of these things are reflected by patients themselves. Next slide, please.

But when we originally made this model, we made it in 2021, even before the diagnosis code was available. And we have since adapted it for the modern era of the pandemic. And I'm going to go over some of the changes that we made because I think that it proves that you can't just make a machine learning model, particularly with a newer condition like long COVID, and just expect it to work as new information comes out and as things change. We need to continually adapt our model to keep up with the science and to keep up with what's going on with patients.

So, as an example, when we first made our model, it really wasn't that certain as to how often people were getting reinfected with COVID. Many of you probably remember that there was even a question as to whether you could get COVID more than once very early in the pandemic. And now, of course, that is not just common knowledge, but many people have had COVID several times. So we needed to have our model understand reinfections, which it did not before. We needed to make our model understand that, particularly now and in the past couple of years, so many people are getting their COVID test at home instead of going to their doctor's office, and that might mean that they're missing a record of their COVID test in their medical record, which means that we need to find alternative ways of identifying that acute COVID period. And then, of course, just the absolute large volume of data that we have now, that we didn't have then, has really made us change our model.

And so, that updated model basically takes what we did before, which is to look at about a... I'm going to say about a two-year period around the acute COVID period. So you can see here, at the top of this figure, we used to have a 320-day data gathering period before a patient's acute COVID, and then a 255-day gathering period after acute COVID, and consider all of that data. But what happened when the patient got a reinfection in 2023, we just didn't even account for that. So we have no data gathering period there. We knew that that was not going to be sufficient.

And so, now, what we do in our new version of the model is that we use overlapping windows of time, where we are checking in every single time window in 100-day period as to whether a patient has had acute COVID, in which case, we black out the data so that we're not capturing information about acute COVID when we're trying to look at long COVID. And we will continue all the way through time in these fixed windows, all the way up until the present day, and patients will end up with scores, that I'll talk about in a minute, that allow our model to say how confident it is that, in that given window, a patient looks like they have long COVID. Next slide.

So what do the model results look like? I find that sometimes this feels very mysterious when we say a model said that a patient is likely to have long COVID, and I think it's instructive to actually open the hood and look at what we would actually see as researchers here. So you can see in this table, and this is all totally fictitious data, but we have four patients: 111, 222, 333, and 444. And each column is one of those overlapping 100-day windows. And you can see we have window one through window five here, and you can imagine that growing all the way up to window 45, and those are fixed in time. For each window, the model is giving a patient a score between zero and one, and that score can be interpreted as the model's confidence that a patient belongs in the "Yep, this looks like long COVID" category.

So we happen to use 0.9, so 90%, as the lowest score at which we are willing to trust the model's decision. And I'll talk about why in just a moment. But you can see here that these scores, highlighted orange in this table, are those that passed that threshold. So for patient 444, their patterns looked very long COVID-isc in window one and window three, and window five. So what we do is use the start of the max... We use the earliest date of the score over 0.9 as the date that we assume that the patient, if they do have long COVID, that that would've been when they actually started seeking care for it. And we use this information to power a lot of our studies. Next slide.

So I think this may be my last slide or coming up on it, but I wanted to just give some caveats about machine learning, because machine learning is not magical and there are flaws to it. One of the things is false positives. So patients that the model is labeling as potentially having long COVID that simply do not have long COVID, and we have no way of knowing as researchers, with data that doesn't enable us to reach out and talk to the patients that we're studying. So one indicator that false positives are happening is the fact that there should be zero long COVID patients before 2020. There just should be. However, that's not the case when we run our model on pre-2020 data, which we did as an experiment. We therefore know that those cases are false positives, and we can use that rate of false positives in the pre-COVID area to estimate how often we're getting it wrong in the post-COVID area.

And that's part of the reason that we selected what seems like a very high threshold, 90% confidence, to say that we trust what the model is saying because that level of confidence seems to minimize the false positives as much as possible. And indeed, when we use that 0.9 threshold in a chart review, meaning that we had a clinician, a long COVID expert, review the charts of randomly selected patients that our model deemed either definitely does not have long COVID, or yes, very probably does, with a sample of 50 charts from a single site, our model had a recall of 0.9 and a precision of 0.7, which are not perfect numbers. Machine learning model will never be perfect, but we feel scientifically justified in using those data to, again, power our studies. Next slide, please.

I think in the interest of time, I am actually going to skip over explaining the slide in detail, but I will note that when we adjust for false positives from that historical control experiment that I mentioned, our adjusted prevalence among N3C patients with COVID for long COVID is 10.4%, which to us, again, looking at the literature and understanding common sense, sounds a whole lot more realistic than that 1.2% that I mentioned before if you only use U09.9. Next slide.

Okay, so here's my last slide. These are the takeaways. Identifying long COVID in the EHR absolutely has trade-offs. You either use U09.9, and maybe feel like you're being a little bit more accurate. That is fewer false positives, but you're going to miss a lot of people. While a machine learning approach has the danger of more false positives, but it also has the advantage of potentially catching more people. Defining the population, and this is really the answer to the question of why does it matter to even do this? Well, it is the first step in using data to answer critical research questions. We first need to know who has or likely has long COVID before we can move on to the next step of answering any research questions.

One of the most exciting use cases for this kind of data-driven research is something called a target trial emulation, which enables us to simulate a randomized control trial for different potential preventatives and treatments before a actual randomized control trial can get off the ground. And that's just one use case for some of these data-driven approaches. And then, what you're about to hear about is how important it is that we not just build and run machine learning models in the data that we have available to us, but that we run them in other environments to try and get a sense of what the validity looks like when we go outside of the data using which the model was built. And that is exactly what Dr. Master is going to talk about. So I think if you do the next slide, I think I'm turning it over to Dr. Master.

Dr. Hiral Master:

Thank you, Dr. Pfaff. Hello, everyone, and thank you for joining the webinar. My name is Hiral, and I'm going to go over how we did this application of the machine learning algorithm that Dr. Pfaff and her team generated in All of Us Research Program. Next slide.

So, again, just to reiterate the goal that Chris mentioned early in his presentation, our primary goal was to validate the machine learning model that Dr. Pfaff and her team had generated in N3C data in All of Us Research Program, primarily because it's a disease-agnostic dataset. And we were really interested in knowing the performance of this machine learning model in identifying and predicting the long COVID cases in a real-world testing, especially in an independent and diverse dataset that we mentioned early in the presentation. Next slide.

So we started this project in May 2022, and I'm going to go... Well, some of the high level, like how did we do this project, what are the key takeaways, and what are the next steps, and how it relates to participants. So I'm going to go on a high level on the outcomes of this project. Next slide.

So, first, when we started this collaboration back in May 2022, in order to maintain transparency, which is a key component, we ensured that this collaborative project information was disseminated publicly as the efforts began. So it was disseminated on All of Us Research Platform, as well as on the RECOVER website. And we provided the information on what this collaboration was meant for, basically, to increase our understanding around this machine learning model for long COVID and how it performs in All of Us Research Program dataset. Next slide.

So now, how did we work on this project? I think that's the number one key question that a lot of people have. Okay, this is amazing. Two big programs are there. How did we actually started the work? Because we wanted to ensure that participants' privacy is maintained, and we ensure that the research is done, so we were very, very careful about how to kickstart and implement this collaboration. So the first step that we did is we did a small multidisciplinary work group, with our primary focus on how we can deploy this model in All of Us Research Program in a secure fashion. Next slide.

So in this, our key people, key players, were NIH RECOVER, N3C, DRC, Data and Research Center, and PCORnet. So the leads from each of these centers came together and would meet on weekly and bi-weekly basis to troubleshoot this algorithm translation, and also if they have questions. So let's just say All of Us team had a question. Then they would reach to N3C, get clarification, and make sure that the algorithm is interoperable, and with a final goal was to make it available to researchers and also to ensure that participants are well aware of these efforts. Next slide.

So, first, what we did is we onboarded the key people on All of Us Researcher Workbench because, as Chris mentioned early in his presentation, the All of Us Research dataset is... The key principle is we want to maintain the data is accessible to widely, but we also want to maintain participants' privacy. So one of the caveat of using All of Us Research Program dataset is that we can only use the row-level participant data on the Researcher Workbench, which is a cloud platform. Downloading this row-level participant data outside of the secured boundary of this Researcher Workbench, the cloud environment, is prohibited. Therefore, the first step was to ensure that we onboard the key personnel to our cloud platform, and they should have access to this controlled tier dataset. And when we do that, we make sure that we provide the information about this research project, like who are the people who are working on this project. So we make all of this information entered in this cloud platform. Next slide.

Then, we also provide information on what is the scientific question, what are the research approaches, and we provide all of these detailed information. And when we create the workspace, in order to get started on working on this dataset, this information is then displayed on Research Project Directory website. So, basically, it's a public website. So anyone who is interested in knowing how All of Us Research Program dataset has been used, they can go back to the Research Project Directory website, which is researchallofus.org. When you go on that website, and then in one of the dropdown icon, you will see Research Project Directory. If you go to Research Project Directory and if you type long COVID, you would be able to access the project that we did. It should come up on its own.

So this project that we did was All of Us. We cover long COVID version six, and this information is available publicly. The scientific question, the methods used, anticipated finding, the key people who worked on this project, everything is available on that Research Project Directory website. And again, that is one of the main thing, main prerequisites of... Even before we start working on the dataset, entering this information and making this all available is number one criteria.

I want to highlight that this information is really dynamic in nature. So what happens is... Let's just say, today, I go on Researcher Workbench on a cloud platform, and if I create a workspace and enter details, given the nature of how research work, sometimes questions evolve, or sometimes our methods change. It's not set in stone. You can definitely go on and update the information and make sure that it's accurate to what the current state of research project is. And what happens is this workspace information that you entered is connected to our website, the public website API. So it automatically translates. So any updates that you make, it is then reflected into the Research Project Directory website in a day. So this was a foundation of setting up a team, making sure that we enter all the details, and then kickstarting the project. Next slide.

Now that we have all this information, now we go into the nitty-gritty or technical stuff. Let's open our Jupyter Notebook, and that's where we can call this row-level dataset. Currently, there are different programming languages which are supported on Researcher Workbench to access and analyze this dataset. One is Python, second is R, then RStudio and SAS. So these are primary program languages that, if you, as a researcher, once you have registered to become an All of Us Researcher Workbench researcher, then you would be able to use any one of these programming languages to then access and analyze the row-level participant data. For the purpose of this project, we used Jupyter Notebook, and we selected Python as a programming language because that was the programming language that our team preferred. Next slide.

So now, one thing which was a benefit of how we were able to translate this machine learning algorithm from N3C to All of Us was both N3C and All of Us had EHR data. And one important thing is that electronic healthcare record data was basically standardized in a similar fashion. So, N3C and All of Us both had standardized EHR data using OMOP. So I would like to acknowledge that yes, we collect EHR data from multiple healthcare provider organization, and standardizing it is very important so that the models and [inaudible 00:47:32] analysis can be interoperable. So having the dataset structured in a similar fashion across both the programs was really of benefit, and it really helped the translation of the machine learning algorithm effectively.

Second is when we were implementing this machine learning algorithm in Researcher Workbench, we realized that the machine learning algorithm was there in PySpark. So there was a need to translate this PySpark and Spark SQL code to Python pandas and Google BigQuery because, again, there is some platform differences. And so, we did initiate these efforts to do the technical translation to ensure that the models are seamlessly translated into the Researcher Workbench, which is the All of Us environment for analysis.

Then, we also conducted a descriptive analysis, where we examined the demographic characteristics of the All of Us Research Program dataset. And we also included those who had less than versus those who have at least 75% long COVID probability, which was computed using machine learning model that was presented by Dr. Pfaff. And next slide.

So the goal, before I jump into the machine learning model, the goal, the work that we did, one of the important factor that we were doing is, as we are working on this analysis, we wanted to make sure that the work that we are doing can be reproducible to other researchers, so they don't have to do this translation piece. So all the code and everything that we have done, we have made it available to All of Us Researcher Workbench researchers. So any researcher who is registered on the Researcher Workbench, they should be able to access the code that we have developed, so they don't have to spend efforts in translating this piece of converting platform transition. That is not needed.

So let's go on how we actually did. So, first of all, in order to implement this machine learning model, collecting cohort is the number one feature. So the model that was implemented was 2022 model. So what we did is we included participants who had at least one COVID diagnosis, and they had at least 145 days after the COVID diagnosis. And that COVID diagnosis can be from condition data, lab measurements data, any of the EHR COVID diagnosis data. We would consider them, and we would include those participants. Next slide.

So, I want to just do a detail, like a quick rundown of how we came to our analytical sample in All of Us Researcher Workbench. So we use a controlled tier dataset that was released to the researchers on June 2022. Our initial starting cohort was any participants who provided EHR data, which was 250,000+ participants. And when we applied the selection criteria, inclusion criteria, we then had 14,000+ participants who had at least a COVID positive diagnosis on condition and lab results. And of that, around 9,000 participants had enough medication and diagnostic data to be included in machine learning models, because we wanted to make sure that they have enough data points that we could run the model, and we could generate the long COVID probability. Next slide.

This is a quick, high-level distribution of what the long COVID probability distribution looked in All of Us analytical dataset. If you see, it's pretty spread distribution. And next slide. What we did is, for the purpose of this presentation, we use 75% threshold to identify long COVID. Now, this threshold can be tailored as per the researcher and their study design. So the cohort that we have made available has 75% threshold. However, if based on your research study, you determine that, "No, we need to be less restrictive. We want to keep 60% threshold," you can definitely alter the thresholds for identifying the long COVID. Right now, we just use 75% threshold for the purpose of this presentation. Next slide.

So based on this, how many participants were then identified as having long COVID? I want to then say that of this 9,000 participants that were COVID positive, we identified 1,800 participants has long COVID, and that is based on the 75% threshold. And then, we also want to highlight that there were U09 diagnosis code that Dr. Pfaff mentioned earlier. Now, since this dataset was active as of June 2022, the number of participants with U09 diagnosis were only 30. But that is expected because the U09 diagnosis code was, at that time, being developed, and it was fairly new. As Chris mentioned, that we are now releasing new dataset in near future, we are hoping that these numbers of U09 diagnosis code would change because the data would be current as of other data times. Next slide.

So I want to give a high-level rundown of what the demographic characteristics look of participants with the long COVID probability based on 75% threshold that I mentioned earlier in my presentation. We saw that one of the key focus of All of Us dataset is basically to enroll participants who have been underrepresented in biomedical research. My demographic characteristics is really focusing on how many participants were either non-white, Hispanic, Latino, female, or were older adults. And if you see on the slide, especially the yellow highlighted table, you can see that around 58% of the participants who had long COVID probability greater than 75% were non-white, 29% were Hispanic or Latino, 59% were female, and around 50% were older adults, defined as age greater than 60 years of age. Now, when you compare these characteristics with participants with long COVID probability less than 75%, you can see that pretty comparable because the difference is less than 10%. We haven't done any statistical testing, but just on a high level, this is what the breakdown of the demographics looks like. Next slide.

This is what the model performance in All of Us look like. Again, as Dr. Pfaff mentioned, the machine learning models are definitely not perfect. There is a room to [inaudible 00:54:52]. So we just have provided... We then explored what the receiving [inaudible 00:54:57] curve, ROC curve, looks like to basically identify the performance of the model. And this slide shows the ROC. Next slide.

And on a high level, I just want to provide a high-level overview. The AUC of the machine learning algorithm in All of Us was 72%. While in N3C, it was 83%. Again, I also want to highlight that the number of cases or the sample size in All of Us was comparatively lower compared to N3C. And that is expected because All of Us is a disease-agnostic cohort. And one of the factor that we use for the COVID diagnosis is EHR data in All of Us.

And it's important to note that the EHR data that was available at the time of analysis only include the healthcare provider organization who are part of All of Us consortium. So it is likely, and we have acknowledged that in limitation, that participants who may have COVID diagnosis positive from a healthcare provider organization who is not part of this All of Us consortium may have been missed. So that is one of the reasons why the sample size in EHR data for the All of Us may be limited is because this EHR data currently that we use with analysis is only comprehensive of healthcare provider organization part of All of Us.

Now, the program is actively expanding on the efforts, and in the future releases, we might also provide EHR data from [inaudible 00:56:37] portal. So what I mean by that is, if a participant connect their own EHR data directly on the portal, we would be able to provide their EHR data as well. What that means is, this might provide the more comprehensive EHR data and would also include EHR data from the healthcare provider organization who are not part of All of Us consortium.

I also want to note that potential long COVID cases were also different between All of Us and N3C, which is, again, expected. And again, the ICD diagnosis code of U09 for long COVID was also a little different. And as I mentioned, one reason why All of Us would have 30 participants that is less than 0.5% is because, when the dataset that was used at the time of analysis was current as of June 2022, and the diagnosis code was fairly new, with the new dataset that might come in the future for the researchers, we might see changes in these numbers. Next slide.

So now, as I said, one of the key thing, not boggling down with a lot of details, is we wanted to make sure that the work that we have done, especially this translation piece, or translating this N3C RECOVER algorithm to All of Us dataset, we wanted to make sure that that translation piece is now available to researchers, so they can build up and do more discoveries and more research because there are a lot of brilliant minds and a lot of people interested in learning and researching on long COVID, and All of Us as a platform where we encourage researchers to expand and work more.

So this code that we have developed, we have made it available in one of our featured workspaces, and it's also available on GitHub. If you're working on Researcher Workbench, I would highly recommend you guys to look at our featured workspace on All of Us. We cover long COVID with this featured workspace. And simply duplicate the workspace, and you would be able to reuse our code and build it from there. You guys don't have to start from scratch. You can directly start from the... We have developed this analytical dataset, where we have implemented this long COVID algorithm. We have generated this long COVID probability. If you guys want to change the threshold of long COVID probability, you can do it. If you guys want to tweak the algorithm, feel free to do it. It is available for all the All of Us Researcher Workbench researchers to then explore and work around. Next slide.

So how can you access and use this in your own workspace? Again, I won't go in detail, but I want to highlight it's very simple. It's just like there's a three-dot hamburger icon. You simply click on that, simply click Duplicate, and once you click Duplicate, that code and everything is then available to your workspace, and you can use that code. So I just wanted to show the highlight of ease of how you can copy or how you can use the code, which is available in this workspace that we have made available to the researchers. Next slide.

As I said, we have created notebooks. Right now, it's in Python. Can you make it in R? Absolutely. Can you do it in other programming language? Absolutely. But right now the code is available in Python programming languages. And we'll be able to go over the key metrics that I've shared, and even deeper dive into the report, as well as the detailed methodology. Next slide.

So, as I said, the first notebook that we make available to researchers is it provides high level, with the proper markdown notation on what this project was, what are the key findings. And we also provide a Python code, actually in the appendix section, so you can replicate any of our findings. If you want to see, "Oh, how did you do demographic characteristics for sex at birth?" you guys... We have a Python code made available [inaudible 01:01:01]. Simply copy paste it and you can duplicate and you can play around with your own.

Second, we also provide all the libraries that are needed to implement the algorithm in Python programming languages. We also show a detailed steps, especially the criteria that we use, like how did we limit at least one COVID diagnosis using condition and lapse data. Then, going onto the next slide of having enough monitoring period, how did we build the code? How did we implement this machine learning algorithm by importing the machine learning model files? Everything has been shown, and it can be replicated. And next slide.

So, again, all other resources that we have, is we have also published a manuscript, and feel free to read that manuscript. And we have also done a presentation, where we do a detailed deep dive, where we show a demonstration on how you can use this work for your own research purposes on All of Us Researcher Workbench. I would really encourage you all to see those two resources to get familiarize yourself. Next slide.

So I want to highlight what were the key outcomes, and I want to just say why this work was very, very important, and how this accelerated our timelines. The first thing that I want to say is it really shortened that timeline for validation. So in traditional research world, it would take years and years of time to basically test, validate, because you have to do data collection, analysis. But in this phase, with the two programs having their own efforts and collaborating and working together, team [inaudible 01:02:47], it really shortened our timeline, not from years, basically to months. We did it in three months, and our goal was to make it available to researchers, so it really helped shorten the timeframe. Next slide.

We also leveraged the resources and made the plot, like the algorithms, interoperable to showcase how the algorithms are interoperable and help advance the long COVID research. Next slide.

This work continues. I will not go in detail, but again, PCORnet is working on doing some of the long COVID computable phenotypes on Researcher Workbench. I'll not go in detail with the numbers, but next slide.

Last but not the least, I want to highlight that All of Us also has other data types besides EHR data. We have genomics data, COPE survey, wearables data. So, again, there are a lot of data types that are available that can help build the model, expand the model further. We are already having some of the ongoing efforts on that. And there's a preprint out over there, and we are working on that. Next slide. I think this is my last slide. So yeah, thank you.

Christine Bevc:

Thank you to our panelists today, and I would like to turn things over to our discussant. Chris, you're going to help kick us off, and then we'll move into our Q&A for all the really great questions that folks have been entering. And thank you to our panelists also for providing written answers to some of those questions as well. So, Chris.

Chris Lunt:

Great, thank you. And I wanted to start off by, and based on some of the questions we've seen coming in the Q&A too, just do a quick recap of the difference between what these different programs are that came together. So, All of Us is a program that is not selecting for any specific disease, but it's just trying to collect information that can help drive science for any group out there that has a specific hypothesis. Whereas N3C was started specifically to study COVID, and then RECOVER was started specifically to study long COVID. And so, they've collected data themselves for the sake of doing that. And so, that was part of the reason that we wanted to be engaged was to go like, "Hey, look, we have a large set of data that you can use to validate the models with an audience that wasn't selected specifically for this." And so, that was part of what brought us all together.

But I think, too, the other advantage for us is that N3C really paved the way for a lot of new approaches to collecting clinical data, which is one of the real challenges here. And I think we saw some questions about that: what about getting records from small practices or high-end practices, and things? And it's something that we struggle with. That the country, in general, doesn't have an easy way to move clinical data around. And there are a lot of different efforts underway to try and make it easier for that data to be able to be something that you can take with you. So if you move from state to state, or you move from provider to provider, that you're able to really build up that continuous record about your specific health.

And so, I wanted to start with covering just a couple areas about the collaboration itself, and then let's get into the specific questions that we've been seeing. And the first one I wanted to ask is, and to both Dr. Master and Dr. Pfaff, what you feel were the biggest barriers to collaboration? What were the things that you had the hardest time getting done, working across these different programs? And I'll ask Dr. Pfaff first.

Dr. Emily Pfaff:

Yeah. I think there are several things that come to mind, but it touches on what you were just saying about how difficult it is to move healthcare data around. And that applied to this collaboration as well, where there are so many really good reasons for the restrictions that we have in place, both in the

All of Us program and in N3C, and in any other system that contains healthcare data. To protect patient data, it is extremely sensitive data. People entrust us with those data, and we owe it to those participants to take care of their data as best we can. However, there is also this other side of that coin, where it can make working together across healthcare data repositories really challenging.

And so, one thing that I felt really great about during this collaboration was the fact that we were able to work around those barriers. It did take longer than it probably would have if there were no restrictions. In fact, I know it did take longer because of that. And there were certainly moments of, I'll call it frustration, where I felt like, "If only I could get my hands on the data that Hiral is using, we could fix this problem a lot faster." But we were able to get it done. And so, I think that collaborations like this, maybe if everyone approaches it with their expectations a little bit lower in terms of time to get things finished, I think that those kinds of barriers are surmountable.

Chris Lunt:

Great. Dr. Master.

Dr. Hiral Master:

Yeah. I think the barriers that Dr. Pfaff mentioned, and that you mentioned, Chris, [inaudible 01:07:55] was most of the things. I would like to highlight, we had the best brains working together. So technical implementation, especially some things like the programming languages, were different, and especially translating from PySpark to Python, making it adaptable, going around the ways, making sure how this data stays, because everything is done on cloud platform. So making sure that it is replicable was one of the challenge, but we were able to get around.

And the second thing was also about the data types. We had EHR data. And again, our EHR data is pretty like structured OMOP data, and they were concerned. "Oh, what if we have clinical notes here? Can we make this more... Can we expand this? Can we validate more?" And I'm like... There were some challenges on that, but again, that's for future research and how we can expand this model further.

Chris Lunt:

Thank you. And now I'm going to get the questions, going to start to transition to where I think a lot of the interest of the audience is. And one is that there's a domain of research and domain of care, and that there are a lot of barriers that are put up between those two very deliberately. And so, it's been a challenge, I know, for a lot of programs to try and understand how do we take the work that we're doing and really make sure that that's delivering improved values for people. And in working on this project, did you see any ways that you thought you could really do that? Did you see a way to be able to translate the work that you were doing into improvements in care?

Dr. Emily Pfaff:

I think, in my own mind, the best opportunity for us to be able to translate this particular work for long COVID is, and something I mentioned at the end of my remarks, which is the idea of simulating or emulating clinical trials. And I think the reason that that is a great opportunity is because it takes data-driven work beyond a lot of the descriptive work that we've been doing for a long time, which is really interesting and important, but doesn't often feel, I think, to patients as if it's as actionable or translatable to their situation.

Whereas I think being able to simulate a target trial using data, and then potentially handing off the results of that target trial to folks who actually had the capacity and expertise to start a real randomized clinical trial, which, of course, is the gold standard, is a very translatable, impactful thing that I think we should be doing more of in the EHR space. But in order to be able to do those target trial emulations, the first step is doing this modeling work that we both talked about. And it is literally impossible to do that more translatable work until we can correctly identify the population under study. And in the case of long COVID, that is particularly tricky in data, which is why we've spent so much time and energy talking through that. But that's really where I see the direction going now.

Chris Lunt:

Dr. Master, anything you wanted to add?

Dr. Hiral Master:

I agree. The clinical trials would be great. I would also encourage people to come on Researcher Workbench and try to see if the... We have this long COVID probability. We have other data types. If there are other ways that they want to explore and have a research question, where they feel that long COVID is critical in their research study design, we have this work done. They can try to incorporate that work into their work and see how it can further enhance their study.

Chris Lunt:

And, Dr. Pfaff, I want to go a little deeper on one thing that you talked about there, and about the fact that the early part of this work is definitional, that it is to try and come up with a common agreement around what are the boundaries, say, of this condition. And recognizing, too, that if you look at things like autism spectrum disorder is something that, 20 years ago, people didn't really know about or talk about, and now it's something that people say, "Oh, it's everywhere." And it's not necessarily that there's more than there was before. It's just people [inaudible 01:11:52] label to be able to put onto a set of behavior that they recognized.

And so, I think it is an important part of this work, is to promote the fact that this is a condition that is understandable, that has some set of commonalities, but it's also a challenge that I think to get buy-in in the broader community on that. And so, I'm curious how you think about that problem. How do you take the work that you've done and promote that in a way that gets into something that becomes a common conversation?

Dr. Emily Pfaff:

Yeah. I try to promote this work and talk about this work in a way that doesn't suggest that we have the right answer. I'd like to think that we have a right answer. But I think that, particularly with the new National Academies long COVID definition that has been making the rounds lately, we've looked at that very carefully, and there's a lot to like about that definition. And it was very carefully constructed, with a lot of input from really important stakeholders. It looks pretty different from our definition, right? And we've started the work, although it's very much underway and not finished.

I'm actually trying to construct a Venn diagram to show what patients meet our machine learning definition that I talked about today, what patients meet the National Academies definition, and which patients are in common between both. And what we're seeing is that there are a lot of patients that are caught by both definitions, but there's also plenty of patients that are only caught by one or the other. And that's not to say that those are the only two out there. I mean, PCORnet has a definition that is also a right answer, and WHO has a definition that is a right answer.

I don't think that we have enough information right now to be able to decide who wins, if that's even an appropriate question to ask. But I think that the union of all of those definitions are... as long as they are evidence-based and well reasoned, and are explainable. And by explainable, I mean that, for example, if it's a machine learning model, that the features that the model is using to make its decisions are not bizarre and don't sound like a clinically reasonable definition. As long as those things are in place, then I think we can use a best-of-breed approach to come up with a suite of different possible definitions to use in this kind of work.

Chris Lunt:

Yeah. And there was a really good opinion piece recently in the New York Times about obesity and about the arguments that go on around how we define obesity. And I think also, then, the problem of the way that people want a binary definition [inaudible 01:14:28] to be on or off, when really, I think that what we really want to do is more stratification to be able to break up the disease into smaller chunks that are understandable, but to have a big tent that brings in a broader set of conditions as well. And so I'm curious, in this work, if you've started to see, do you think there's a stratifiable group of people here that'd be worth dividing [inaudible 01:14:45], "Oh, there's this version of long COVID, and this version of long COVID, and that those are things we should think about and try and research and study differently"?

Dr. Emily Pfaff:

There has been work done by all of the RECOVER EHR cohorts along those lines. We refer to that as sub-phenotyping studies, where we do attempt to put people into a category that best fits them. As you can imagine, that's not necessarily easy because long COVID affects so many different body systems that it's not necessarily mutually exclusive, right? So if we have a cardiopulmonary category, that doesn't mean that somebody's not also experiencing neurological symptoms.

And so, some work, and I would be happy to follow up with some citations of work that's been done in this space, have tried their best to pick the most winning category for different groups of participants, and have found different patterns, such as cardiopulmonary clusters of patients within the long COVID, patients experiencing neurological symptoms that have long COVID. There's even a gastrointestinal cluster. So, yes, there are absolutely sub-phenotypes that can be explored further and may need different treatments and different preventatives.

Chris Lunt:

And along those lines, and I think I saw you'd answer a question in the chat about post-viral syndrome, which... So there's an even bigger tent of adjacent conditions that we've seen for a long time. And I'm just curious, what opportunity you see from the work that's being done here to try and also help to chase down that longer problem that we've had dealing with all of the post-viral syndromes?

Dr. Emily Pfaff:

Yeah. There is an ICD-10 code, B94.8, which is a catchall code for post-viral sequelae, and that's been around for a really long time. And when we started doing this work on long COVID, we actually wanted to look at what the usage of that was like, even in the pre-COVID period, and the numbers were very, very small.

Once long COVID really came to the forefront, and before there was a specific diagnosis code available for long COVID, all of a sudden, we saw usage of that B94.8 code ticking up and up and up, showing that there was, at least, increasing recognition of this idea of post-viral illness.

And what I think would be very interesting to do as future work is looking farther back in time. We have records that go back quite a ways into the end of the '90s, and actually try to apply some of these models that are identifying long COVID patients to see if we can identify patterns in post-viral illness patients in the before times. I think that would be really, really interesting work, and would maybe bring to the forefront some of this post-viral illness prevalence that really hasn't been talked about enough in the past.

Chris Lunt:

I think it was one of the most exciting studies I remember reading three years ago, I think, is when it came out, and this was out of the Million Veterans Program that's run by the VA, showing the association between multiple sclerosis and Epstein-Barr virus, and that's something I'd never had any suspicion that that could potentially be a post-viral syndrome.

Okay, one of the other questions that has been coming up in the chat has been about insurance. And so, I'm going to try and see [inaudible 01:18:00] that a little bit into about the willingness of insurance companies to pay for this, and how does this work support trying to get the payment part of care to be a little more equitable as well.

Dr. Emily Pfaff:

So I have to admit that I am not a health economics person, and so I don't want to speak out of turn. What I will say is that we did see that pattern of slow uptake of the new code when it became available, that U09.9 code. And I think I saw someone else in the Q&A reference the fact that there may not be enough provider education to tell providers that, "Hey, there's this code for this, and you can enter it into the chart," as well as that there may just not be enough promulgation of the guidance from CDC that shows how to use that code and when it's appropriate to use that code.

And while I admit that I do not have an answer to that, and I don't know what the implications are for reimbursement insurance, I will say that we have seen that gap between, there's guidance for using this code and actual use of the code. And I certainly think it is fair for patients to bring up with providers that, "There is a code for long COVID, and is it appropriate to be on my record?"

Christine Bevc:

Thanks.

Chris Lunt:

That's great. Christine, I should probably turn over to you now, and you can use that one-on-one direct Q&A.

Christine Bevc:

Yeah, we've got to-

Chris Lunt:

[inaudible 01:19:26], I'll talk to these two all day.

Christine Bevc:

Oh, I know you will. Yeah, and that's great, because these are the questions we've seen coming in through our Q&A, and that question about the implications of insurance, using the code, the challenges there. Real quick question, which I know we answered in there, but is All of Us still admitting participants, and what about those in Canada?

Chris Lunt:

Yes, All of Us is still letting in participants. We are going through a platform change this December, so it'll be a week where we're not available. But it is for US resident and people living in the US only.

Christine Bevc:

Okay. All right. And then we'll include a link of how individuals can learn more about participating in the program. These, to the group, are also about for the future of including participants under the age of 18. Are there efforts headed in that direction?

Chris Lunt:

I can give, and I'll give a quick answer, is that All of Us started pediatric enrollment for people ages zero to six this year, but on a small scale. And our interest is there to expand that to go to larger ages and a larger group, and that's dependent on our budget, and that's depending on what Congress does since we're under a continuing resolution at the moment, and Congress has to decide by December 20th what the next steps are for funding government programs.

Christine Bevc:

All right.

Chris Lunt:

And sorry, and I can answer quickly for N3C as well. N3C does collect pediatric data, but Dr. Pfaff told me that this model does not include children. It's not built for children, and this only works for adults.

Christine Bevc:

All right. So for the work that's ongoing, we've talked about the catalyst being able to access some of the workbenches there. Can non-researchers join All of Us to be able to view some of this ongoing research that's happening?

Chris Lunt:

Yeah. If you're not a researcher, but you're interested in what's going on, there is a research highlights newsletter that we put out regularly, and you can find that if you go to researchallofus.org. You can find a link, and I put a link in the Q&A as well.

Christine Bevc:

Okay, great. Thank you. And then, more generally, I just wanted to take us back a little bit more about, again, highlighting, if each of you can speak to how this program is really working to improve the lives of long COVID patients. If you had to sum it up, and I'll start with you, Chris, then I'll go to Dr. Pfaff, and then Dr. Master.

Chris Lunt:

Yeah. I think we touched on this a little bit, the problem of translation. We're a research program, so we don't directly engage with care organizations. We don't directly try and create changes in standards of care. We're not funded to do so, or empowered to do so, but we've been in conversations with the National Center for the Advancement of Translational Sciences, or NCATS, which is an NIH group that is responsible for doing that, to talk about how we can accelerate the process of, if we discover something, how do we get that to show up in terms of improvements in care.

That process, in my mind, takes way too long right now, and it is something I'm very interested in seeing what we can do to accelerate that. And we're on one end. We're on that front end of discovery, and we're trying to see what we can do to more quickly do that. And as Dr. Pfaff talked about, too, I think this is where clinical trials and the capacity to accelerate the integration with clinical trials, which will be important. And the current director of the NIH, this is one of her big priorities as well.

Dr. Emily Pfaff:

I guess my response to that question would be that I want to see us in the data space doing more that directly impacts patients. To Chris' point, that can be difficult. Obviously, I'm not a provider, and nor are most of my colleagues in data science world, but that doesn't mean that we don't want to have an impact on patient lives. And sometimes, the discoveries that we make in data feel important scientifically, but are not things that patients can reach out and touch and feel like are making positive changes in their health. And not to continue to repeat myself, but I do think that our future involvement in clinical trials is probably our best way, as data scientists, to contribute to findings and discoveries that actually will feel impactful to patients in a way that maybe some of the other data-driven work has not.

Chris Lunt:

I want to add one quick thing to that too. I think one thing that we can do to help is specifically to look for ways to reduce the cost of diagnosis. And so, finding new biomarkers is an example of a really important way to really help accelerate things like this. If there's a very easy way to test for something that then you know is treatable, that's helpful.

This is an area, too, where ARPA-H, which is a new government agency responsible for trying to try out bigger ideas that can change health, is doing work on trying to develop new biomarkers. And so programs like ours can help support that by the data that we provide to show how people's health is measurable in a way that makes it easier to identify when somebody has a specific treatable condition.

Dr. Hiral Master:

Yeah. I would just like to add one thing, echoing Chris and Dr. Pfaff's comments, is like use these resources, the big program resources, to basically generate your hypothesis. And again, translate that to your lab or start a small clinical trial, see if you can get data, and try to test your hypothesis. Again, I know the mechanism that Chris said. It's going to take time. But I think, though we don't have direct [inaudible 01:24:56] affecting the quality of life, but I think this work provides substantial evidence, which helps in generating the hypothesis, and then, as the researchers and as the scientists, take a next step and try doing small exploratory work.

Christine Bevc:

All right. And with that, we could go for another hour, probably two hours, with all the great questions that everyone's been submitting, but we are reaching our time for today. So I just want to say thank you so much to our presenters, and thank you to our audience for attending today and really engaging with the Q&A.

As a reminder, the recording of today's seminar is going to be available on recovercovid.org within a few weeks. We're also going to be posting an accompanying Q&A document. That's going to have the responses to the questions that we receive today, including quite a number of those that we didn't have a time to address during today's session.

But before we conclude, today is all about data and availability and what's going out there, but researchers, both within and beyond the RECOVER Initiative, can also now apply to use some of that RECOVER data that was mentioned today. That includes the study data and even biosamples for ancillary studies. And so this is now open, in addition to the All of Us portal and information there. But RECOVER ancillary studies are now available to allow researchers, if they're interested in conducting independent research projects, to help contribute to the search and find answers to the important questions around long COVID.

There's resources available online. This includes the data from our three RECOVER cohort studies, so adults, including pregnant adults, pediatric, and our [inaudible 01:26:55] study, as well as some of the biospecimens that have been collected from our study participants. So if you're a researcher, if you're independent researcher interested in submitting an ancillary study proposal and would like to receive approval to access the data and the samples, you do need to have some independent funding support to conduct your study. But you can find out more on how to apply on recovercovid.org/ancillary. You can read about the submission process and how to access those resources.

So we've thrown a lot at you today. Again, that's going to be posted with the links on recovercovid.org in the next coming weeks. But with that, we hope that you will join us again later this month. We actually have another seminar before the end of this month. That registration, as just like this one, is posted on the RECOVER website. It's going to be open and shortly, as well as those that are going to be further into November. So you can check the R3 page for updates.

And then, to close us out today, you're also going to see a short survey that's going to come up on your screen. It's going to ask for feedback about your thoughts for today's seminar, as well as how we can improve for future seminars and any topics there. So take a moment to complete that, and we appreciate that. This concludes today's R3 Seminar. Thank you to our panelists, and thank you to you, our attendees. Have a great day.