Transcript

Christine Bevc:

Hello and welcome. I'm Christine Bevc, task lead for RECOVER at the Administrative Coordinating Center and moderator for today's webinar. I'd like to welcome everybody to today's RECOVER Research Review or R3 webinar. The overarching goal of the R3 webinar series is to catalyze the formation of a scientific stakeholder community within and beyond the RECOVER consortium. Fostering a shared understanding of the state of science and providing an educational resource for both RECOVER investigators and the broader scientific community of clinicians, patients, and stakeholders.

I want to start by thanking everyone who submitted questions in advance. As Shane mentioned during today's webinar, please use the Q&A feature in your Zoom windows to submit your questions. After the presentation, our presenters will answer as many questions as possible and provide responses in real time within the Q&A. Please note that we will not be answering any questions about clinical care and the data and results in today's presentations are all based on published or pre-print materials. NFAQ document for this webinar will be posted along with the recording of the webinar on recovercovid.org. Today's webinar is the first in our exciting new series of research in action. Today's session is entitled, RECOVER in Action Characterization of PASC Among Adults, EHR Insights.

Our presenters will be addressing the question, what is the clinical spectrum of PASC including subphenotypes? We'll be using EHR real-world data and data science. If you haven't already, please remember to sign up on our website to receive future announcements and updates on the series. So to get us started, just as a disclaimer posted here says, "RECOVER continues to grow, and we want to remind you that the information presented in the seminar is intended to stimulate collaborative dialogue amongst the RECOVER scientific community, as well as study participants and other interested parties. The information may be recently published or about to publish and as such is potentially subject to change. In addition, none of the information here should be interpreted as medical advice."

And without further ado, I'm excited to welcome our four RECOVER investigators joining us today. It's my pleasure to introduce Dr. Tom Carton, Dr. Emily Pfaff, Dr. Melissa Haendel, and Dr. Fei Wang. They are joined by our discussion, Dr. Megan Fitzgerald, who will kick off our discussion and response portion of our webinar. We're excited to introduce our first presenter today, Tom Carton. He is the Chief Data Officer for the Louisiana Public Health Institute and also serves as the Executive Director of the Greater New Orleans Health Information Exchange, which coordinates care and exchange of information across dozens of primary care, behavioral health and hospital systems in the Greater New Orleans area in the state of Louisiana.

Dr. Carton has held multiple leadership positions within the National Patient-Centered Clinical Research Network, including Chair of the PCORnet Steering Committee. Dr. Carton today is going to be providing an introduction and overview of the RECOVER electronic health record cohort study. He'll be sharing details on the PCORnet analysis and the query-based approach using the computable phenotype definitions. Next, we'll be hearing from Dr. Emily Pfaff who will introduce us to the National COVID Cohort Collaborative, also known as N3C. She'll be speaking to how EHR real-world data is being leveraged using two different methods to identify subphenotypes based on the symptom presentation.

Dr. Pfaff is an assistant professor in the Department of Medicine at UNC's Chapel Hill School of Medicine and co-director of the Informatics and Data Science Core at UNC's CTSA. Dr. Pfaff is also one of the leads of the National COVID Cohort Collaborative, N3C, which is one of the largest harmonized clinical data sets in the U.S. and multiple PI of the EHR and real-world data component of the NIH RECOVER Initiative. Our third presenter, Dr. Haendel, will take us deeper into the N3C data and the clustering of patient symptoms uncovered by the dual method approach with semantic similarity and topic modeling. Dr. Haendel is the Chief Research Informatics Officer and Marsico Chair in Data Science at the University of Colorado Anschutz Medical Campus.

She also serves as Director of the Center for Data to Health where she carries out her vision to weave together healthcare systems, basic science research, and patient-generated data through the development of data integration technologies and innovative data capture strategies. And we'll round out our presentations with Dr. Fei Wang and the presentation on topic modeling using the PCORnet data to reveal PASC subphenotyping. Dr. Wang is an Associate Professor of Health Informatics in the Department of Population Health Sciences and Founding Director of Institute of AI for Digital Health at Weill Cornell Medicine. His research interests are grounded in machine learning and artificial intelligence for medicine. We'll also be joined by Dr. Megan Fitzgerald who will lead off our discussion with a short series of questions.

Dr. Fitzgerald joins us today as a patient representative for RECOVER, having become ill with COVID in March of 2020 and subsequently developing long-COVID. She also serves as the co-lead on the study of COVID reinfections with the patient-led research collaborative on COVID-19 and holds a PhD in neuroscience and neurology with a background in human neuroanatomy, developmental neuroscience, and STEM research. We have a great panel as you can tell. So without further ado, I'm going to go ahead and turn things over to Dr. Carton to get started. And then following the questions from our discussion, we'll go ahead and open the floor to questions. All right, Dr. Carton.

Thomas Carton:

Thank you, Christine, very much for the introduction. Thanks to everyone on the panel and thanks for everyone on the webinar. We're excited to be presenting to you guys today. You can go to the next slide please.

Okay, great. You can go to the next slide as well. So I'm going to frame the discussion within the RECOVER community. I'll introduce the PCORnet work and then I'll talk a little bit about our PCORnet rules-based computable phenotypes before handing it over to Emily. And then Emily and Melissa will go, and then Fei. And then we'll go to Megan for the discussing of questions. So this slide essentially, basically places us within the RECOVER consortium. You'll notice the executive committee, steering committee at the top, the various operation groups and committees in the middle, and then the hub sites from adult to pregnancy to pediatric to autopsy.

And then three EHR hub sites, two PCORnet sites, one adult, one pediatric. Fei and I are both parts of the adult sites. And then the N3C consortium is the third EHR hub. And so we are presenting really on behalf of the EHR cohorts within RECOVER. Next slide. We're also contributing on data science and real-world data. So I wanted to just frame this slide very briefly in terms of our pending presentations. And so we conceptualize real-world data and EHR data generally by scale, geography, target, flexibility and efficiency. And those attributes allow us to create subphenotypes, advanced machine learning methods, including simulated clinical trials, leverage previous machine learning models for high throughput analysis, and emphasize the importance of data science and AI as approaches for inquiry. So please keep these attributes and applications in mind as we go through the presentations and the results today. Next slide please.

So now I'll briefly frame the PCORnet piece of the EHR cohorts. PCORnet is the National Patient-Centered Clinical Research Network. There are 65 sites, all of them are outlined here. The 41 that are highlighted in the lighter blue are the sites that are contributing data to the RECOVER EHR cohort. And then there are some characteristics of the data both in terms of size and structure and origination that are given in the box over here to the right. Next slide please. And then within that adult PCORnet team, we're organized in four work streams. We have an epidemiology health services research team. We also have a team that responds to queries from NIH patient groups, clinical science core and others. We're going to focus today on the two that are in the middle, the machine learning and artificial intelligence group, which Fei leads and will present on later in this webinar.

And then the screening computable phenotypes, which are rules-based which I'll be presenting on today whereby we are create rules-based definitions using EHR data for both incidents and worsening of existing conditions as they relate to PASC. Next slide. Okay, so my piece is a little unique via and viz the other presenters who are going to be speaking of machine learning methods, topic modeling, clustering, subphenotyping. I'm going to speak about computable phenotypes that are built with rules that are contributed to by both data science and clinician and patient teams and use EHR data and standardized definitions. And so this process works through how we go about doing that. We generally start with a first question around a organ system or a disease characteristic and we produce a list of relevant lab diagnostic and medication codes to inform the definitions of the computable phenotype.

We then meet with clinician groups, data scientists, patient groups to put forth specifications for each computable phenotype. Within that, we run queries to obtain incidents and prevalence rates. We share those back with the clinician team for face validity and we basically agree upon a draft version of that computable phenotype. In order to evaluate the sensitivity and specificity of that phenotype, we do chart reviews, hundreds per each computable phenotype at sites that are contributing data to the PCORnet EHR cohort. We receive the results of those reviews. We evaluate those for both sensitivity, specificity, face validity. We make any adjustments as required. And then we make final determination and definition of the computable phenotype. And then we disseminate that. We disseminate both the definition, the results from the chart review evaluation, and a basic descriptive profile of each of the cohorts. Next slide.

And so this lays out the phenotype work that we are doing. The phenotypes are listed down on the first column on the left, everything from type 1 and 2 diabetes, cardiology. We have pulmonology, neurology phenotype, pregnancy phenotype that is being developed with the pregnancy cohort. We have a phenotype that encompasses Any-PASC or all derivations of PASC. And then a phenotype on lupus. And then the other three columns, both development resolution, REDCap. And then chart reviews outlines our current state of each in this process. Next slide. So I have two examples that I'm going to give very briefly. And as was mentioned, these are preliminary results. They have not been yet sent to peer review and we are reporting results from limited data marts which are given down below, which are basically health systems that are contributing the data. Because these are the sites that are doing the chart reviews for these corresponding computable phenotypes.

So the pulmonology phenotype covers both COPD and asthma. Here you can see the definitions for both incidents and exacerbation and the data marts that were used to run the preliminary analysis and are conducting the chart review. Next slide. These are the results of the preliminary analysis. We have incidents on the left and we have exacerbation on the right. We have rates of per 100 patients across both COVID positive and COVID negative. For incident conditions, these are conditions that were not present before patients got COVID-19. And for worsening, they were patients that had those conditions pre-COVID and then they were exacerbated post. At the bottom of each of these tables, you have a number that gets that burden. This is excess burden amongst COVID positive versus COVID negative patients.

And here you can see that the incidence rate of COPD and asthma across those five data marts was about 5.4% higher for COVID positive than COVID negative patients. I'll also point out that incidence rates were higher amongst hospitalized patients that had a more severe COVID index infection. We have similar trends within hospitalized versus non hospitalized for worsening, but worsening of existing COPD and asthma conditions we're only half a percentage point higher for COVID positive than COVID negative patients. Next slide. This slide outlines our coronary heart disease computable phenotype, which is our cardiology phenotype. Again, incidents and exacerbation definitions given above. Data marts used are different in that this is where the preliminary analysis is coming from here and where the chart reviews are being deployed.

Each of those health systems are in New York City and are members of the INSIGHT Clinical Research Network. Next slide. And this slide is set up similarly, in that we are looking at incidents on the left and exacerbation of CHD on the right. The top tables for each describe all patients. So about a 0.29% excess burden among COVID positive versus negative for incidents CHD, whereas a 3.2% excess burden for COVID positive for worsening. And you see similar trends across that we saw within the COPD and asthma as well. Next slide. I want to just speak briefly to the way that we have constructed our Any-PASC definition because it really connects to the other presenters here today. We started with 137 condition categories. We ran machine learning and artificial intelligence models as well as looked at the literature.

Within that, we identified 44 conditions of interest and they're outlined in those various colors on the chart in the middle. Within those, we looked at varying degrees of incidents and those significance, reviewed this with clinician groups and then refined this list to a narrower list of 25 conditions that are given at the bottom that contribute to our overarching definition of Any-PASC. Next slide. I'll speak briefly to the chart review process so that folks can get a sense for how these computable phenotypes are being evaluated. We start with an initial survey that is developed off of the definitions for both exacerbation and incidents. We again review that with our clinician groups. Then we create a REDCap instance, which is essentially a data collection form to investigate both the incident and the exacerbation definitions.

We generate patient lists from the data that we have contributed by our sites. We share that information with abstracters at the sites who are trained in this process and then are deployed to evaluate the clinical records and answer a series of questions that allow our data science teams to evaluate the sensitivity and the specificity of each of the computable phenotypes that we've created. Next slide. And then finally, this outlines the reporting format that we have in place for each of the computable phenotypes. So we run a query on the clinical data where we look at both the incidents and the exacerbation rates similar to what we just looked at today, except it would be across all data contributing sites within RECOVER. So about 35 adult sites.

We interrogate those rates by demographics and by comorbidities and also by treatment during their COVID-19 index case. We do variety of analysis on the chart reviews including looking at positive predictive value, analysis of the comments and the notes. We compare ways that patients entered the cohort. There are different cut points based upon different clinical elements. We look at any other possible data that is collected through those REDCaps. These reports are then provided to our clinical teams, which really take the lead in the dissemination of this information through both manuscripts and presentations and so forth. Next slide please. Okay, well, thank you. I am going to now turn it over to Emily Pfaff. Emily.

Emily Pfaff:

Good afternoon everyone. I am looking forward to talking about some of the work that the N3C group has been doing along the lines that Tom was just describing. Next slide please. So as background, N3C is the National COVID Cohort Collaborative. Like PCORnet and PEDSnet, we are part of the RECOVER EHR group. We are the largest national public HIPPA limited data set in U.S. history. So that is EHR data or electronic health record that is anonymized but available for research to researchers across the country. We have over two billion clinical observations now from many, many sites across the country that have opted to share their electronic health records for COVID research purposes. We have a representative cohort that is incredibly diverse as far as race, ethnicity, gender, geography, socioeconomic status and health background.

We have harmonized the data so that data that's coming from one institution can be matched up with and merged with data from another institution. Even though they may not look the same at the source, that makes it much easier and much more accurate to do cross-institutional research. We have data links to outside data sets that can add richness to electronic health records such as vaccine data, data from claims, insurance claims like Medicare and Medicaid, environmental data, social determinants of health and other things. We can use N3C for projects like RECOVER because it is a very rich data set of COVID patients, as well as matched controls, and that enables us to use it for researching long-COVID as well. Next slide please. So as Tom was discussing in the PCORnet context, N3C has also developed a computable phenotype or definition, electronic definition for long-COVID.

We have done that in a couple of different ways and we've actually presented about that on a prior R3 seminar. And so I'm not going to be focusing on that for this talk. Rather, what we're going to be talking on today is subphenotyping, and that means that let's say we consider long-COVID to be a big bucket that we can categorize patients in. So we have all of the patients in N3C, we come up with a definition or computable phenotype for long-COVID, and we have patients that are in that bucket. But we know that from evidence that has come out in the literature and from patient experiences that everyone in that bucket is not the same and every presentation of long-COVID is not the same. And because of that, we think that it's highly likely that different treatments are necessary for different presentations of long-COVID.

Some patients may have only respiratory symptoms of long-COVID. Some patients may have only neurological symptoms and everything in between. And so in order to try to figure out what those different types may be, we are engaging in a number of different strategies to do what's called subphenotyping or defining these different categories of long-COVID as opposed to considering it all one thing. We're going to be talking about four different strategies that N3C has used thus far to do this subphenotyping activity. I'm going to be talking about two of them, the top two here, diagnoses co-occurring with U09.9 and clustering our machine-learning-derived features. Those things will make more sense when I describe them. And then I'm going to hand it off to Melissa Haendel to discuss semantic similarity and topic modeling. Next slide please.

So I'm going to start with the first approach, which is diagnoses co-occurring with U09.9. For those of you who are not clinicians or haven't worked with EHR data, U09.9 is what's called the diagnosis code. It is the code that is entered when a patient comes into a clinic and is diagnosed with long-COVID, that that disease has a specific code that is standard across the U.S., that when it's entered in the medical record, that is a, "official" diagnosis for long-COVID, and the code is U09.9. The catch is that U09.9 was introduced for use in the U.S. on October 1st, 2021, which was well over a year after patients started recognizing long-COVID symptoms. And so what ended up happening, as you can see in this figure here is use of U09.9 is shown in pink here.

You can see that the use of that code did not really take off until almost 2022 and beyond. And because of that, anyone who was seeking care for long-COVID between 2020 and the release of that code may not have had the opportunity to get that code recorded in their record. And so that's a limitation that we have to keep in mind as we as data scientists want to use the U09.9 code as a tool to help us identify patients with long-COVID. However, despite that limitation, it is a useful tool even though it's not the cure-all here. And so with a growing number of patients that do have this code in their record, we have been able to construct clusters of other diagnoses that tend to co-occur with both U09.9, so meaning that they are issued in the same care encounter as U09.9 and also co-occur with each other.

And then if you go to the next slide, I can talk about what that looks like. So what you're seeing here only applies to patients under 21 years of age. In the subsequent slides you'll see the same kind of diagram for patients of different age ranges. But what you're looking at are clusters of diagnoses that both again occur with U09.9 in the same patient in the same care encounter and also co-occur with each other. And these different diagnoses tend to cluster together in different categories. The categories aren't completely clean. So when we say gastrointestinal cluster in the purple on the bottom, obviously not everything in that cluster is specifically a gastrointestinal symptom such as fever. But we have sort of applied themes to these different clusters to describe what the majority of the concepts within seem to be. So when we look at these co-occurring conditions for pediatric patients, we see distinct neurological and cardiopulmonary clusters.

We see an upper respiratory cluster and we see a gastrointestinal cluster. And I want to point out as I move forward into the other age ranges that we looked at, that the upper respiratory cluster and the gastro cluster are actually unique in our work to pediatric patients. Now that's not to suggest that adult patients with long-COVID don't have gastrointestinal symptoms and don't have upper respiratory symptoms, but it was particularly apparent in the pediatric population enough to make those diagnoses their own cluster. Next slide please. So this is patients from 21 to 45 years of age and you'll see that those two clusters that are unique to the pediatric population have disappeared and some of the diagnoses that were within those larger clusters have now sort of scattered themselves within other clusters.

You now see that there is a neurological cluster and that's one feature that is apparent in there is chronic fatigue syndrome and you'll see that throughout all of the age ranges as part of the neurological cluster. Makes sense that that continues to co-occur with U09.9 as well as many other of the neurological symptoms that we often ascribe to long-COVID. The cardiopulmonary cluster remains. We now introduce what we're calling the comorbidity cluster, which are additional conditions that patients may have that may not be directly related to long-COVID but may just sort of come along with long-COVID, just conditions that patients have as they continue to age and may affect the presentation of long-COVID. Next slide please.

In the interest of time, I'm not going to go through these in detail, but as I go through the ages, you'll see here that the neurological cluster is increasing in size as well as the comorbidity cluster. Because as patients age, they're more likely to have comorbidities. Next slide. And then finally, in patients who are 65 years and older, the comorbidity cluster is quite a bit larger, but we still see that neurological cluster and cardiopulmonary cluster that remain. Next slide. So in summary, again, we have identified the diagnosis that are most commonly occurring with U09.9 and have clustered them based on their rate of co-occurrence with each other. And those clusters include the ones that are listed here. Now there are limitations to this approach. The U09.9 population is skewed demographically and temporarily.

So I mentioned that early pandemic patients may not have had the opportunity to get this code. We do see that the U09.9 population tends to be more female, more white, and more non-Hispanic than what we suspect the overall long-COVID population looks like. So there is a skew there that we need to keep in mind. And as I mentioned, those clusters, the names of them can be a bit of an oversimplification, in that there are diagnoses within them that don't quite match the label. Next slide. So the other approach that I'm going to talk through are machine-learning-derived features. Again, our computable phenotype for long-COVID not regarding some phenotypes is machine learning based, meaning that we look at electronic health records for patients who we know have been diagnosed with long-COVID.

We learn patterns of clinical features from their records and then apply those patterns to all other patients in our data set so that we can see if anyone else follows those patterns. And then we can make an assumption about each patient's risk of being a long-COVID patient or not. Next slide. Next slide please. Thank you. So when we run machine learning models, there is a way for us to determine what features the machine found most important in making its decision as to whether a patient is likely to have long-COVID or not likely to have long-COVID. Each feature that is used by the model, and you can see some example features here on the left like age, dyspnea or shortness of breath, fatigue, et cetera. So each of those features gets a score for its importance to the model.

And an important feature could be highly predictive that a patient may have PASC or it could be highly predictive that a patient does not have PASC. So for our subphenotyping effort, we took the most important

features for our model and we aggregated them into categories based on similarity and input from clinicians, and that resulted in some category labels that we could then reapply to patients for use and analysis. Next slide. These categories ended up being this list that you see here on this slide with features in the category of pain and inflammation, pulmonary and respiratory, and neurological and nervous system being the most apparent. But you can see that there are many others that decrease in prevalence as we go along and that these features continue to show up across the eras of COVID. Next slide please.

So this approach, unlike the U09.9 approach, uses our machine learning model and categorizes patients based on the features of their record that contributed to their being labeled by our machine learning model as potential PASC or not potential PASC. The top 10 categories that you see here are those that are deemed most frequent or most prevalent within our likely PASC population. And you'll see in this list many things that have been well reported in the literature as symptoms of long-COVID, including things like sleep issues and gastrointestinal features, fatigue, et cetera. There are limitations to this approach as well. However, machine learning is not a black box necessarily, but it can be difficult for humans to explain all of the logic that a machine uses in making its decisions.

And so sometimes that makes it harder to explain the logic that is behind what you see here. We can make our best guesses. And so for example, you'll see that there's a category of anesthesia, which doesn't seem like a very obvious feature for long-COVID, but we believe that that anesthesia category is likely a proxy for people who were hospitalized or intubated with their acute COVID infection. So sometimes you have to do a little bit of extra interpretation with machine learning models. The other caveat here is that temporality is not considered here. So we look at all of the symptoms that patients have reported over the entire course of their long-COVID and acute COVID. And because we don't actually look at the progression of symptoms saying that first patients had pulmonary symptoms and then that graduated to neurological symptoms, et cetera.

We're just treating everything as having happened at the same time. And that is future work that we would like to get into is actually taking more of a progressive temporal approach to this so we can see if symptoms evolve over time. Next slide. I'm going to hand it off to Melissa to talk about our other two subphenotyping methods.

Melissa Haendel:

Thanks very much. It's really great pleasure to be here today. And so as Emily suggested, we're going to talk about some different approaches that we've taken to identify subphenotypes of PASC. I'll remind folks that one of the goals of all the phenotyping exercises that we're describing today is really to identify subpopulations of PASC patients so that we might better diagnose them, better target them for specific research studies, and better reveal underlying mechanisms of the symptoms and features that they exhibit so that we might hopefully ameliorate those symptoms. So these are all approaches that are complimentary and synergistic to hopefully get

us where we want to go with the entirety of the RECOVER research program in aiding patients conditions. So next slide please.

So this approach is actually something that has been really exciting for our team that we've really envisioned for a very long time and we've finally been able to execute this work. I especially want to thank Justin Reese, Peter Robinson and Charisse Madlock-Brown who really led a lot of this work. So basically what this approach does is it tries to understand the characteristics of each individual patient and understand how these patients are similar or different to one another and then use that similarities and differences to reveal underlying clusters. So for example, over there on the left we have a patient that has a OMOP concept. The OMOP is the common data model and the concept identifier that you see there is referring to a respiratory finding.

Similarly, we have acute kidney injury and we have chilblains. Another patient, as shown on the right, which has this different set of features, here for example, cough, transient renal failure, clearing throat or hawking, and chilblains. You can see that the chilblains is actually an identical match, whereas we have some kind of conceptually partial matches between a respiratory finding and cough, acute kidney injury and a transient renal failure that a human would know are somewhat related. And in the context of this clustering approach, what we've done is create a computable way of relating these non-exact relationships between things like respiratory finding and a cough to understand the similarity and differences between individual patients. Next slide please.

So in order to do that, we've leveraged a resource called the human phenotype ontology. And an ontology is a little bit different than more traditional clinical terminology such as the International Classification of Diseases or ICD or LOINC, the terminology that's utilized for encoding laboratories. And so in this case, what we're really doing is describing patients individual phenotypic features similar to we would any biological subject. And so for example, we have different branches of this ontology such as constitutional symptoms that would include fatigue and pain, abnormality of the nervous system physiology, that might include headaches, which you can see are both constitutional symptoms and abnormal nervous system physiologies and so on.

And so these phenotypic features are represented as a graph structure as opposed to a single taxonomy like Linnaean taxonomy might be, which is how the ICD classification is structured. What this allows us to do is really sort of develop more sophisticated computational algorithms for understanding those similarities and differences based upon the similarity to the individual features in the graph. So for example, headache is actually closely related to chest pain in the context of constitutional symptoms but less related to hypoxemia over there on the right. And so that graph structure allows us to technically calculate the similarities and differences based on that general versus specific representation shown over on the right. I'll show you how that works on the next slide please.

What we've done in this case is leveraging a number of different resources including a library developed by a graduate student, Dr. Tiffany Callahan, who has recently graduated, called OMOP2OBO. We have actually

taken many thousands of patient features represented with ICD codes and LOINC codes as well in the context of the EHR record and transformed those or mapped those into human phenotype ontology terms so that we could leverage this specialized semantic similarity approach. And so what that result looks like is over there on the left. You see patient number one who has features such as fatigue, chest pain, headache, depression, visual hallucinations, and then we're comparing that patient to a similar patient.

In this case, patient number two who has some similar and some different features such as asthenia, which is a non-exact match to fatigue through the most common ancestor term, which I showed on the prior screen as a constitutional symptom in common. Some of the features are exactly identical such as chest pain, headache, and depression. And others are different, again, such as hallucinations where we have visual hallucinations for patient number one and auditory hallucinations for patient number two. So this is what two quite similar patients look like and the number of features on any given patient can range from several to hundreds. And so the algorithm calculates the degree of similarity based upon these matches. Over there on the right, we see two patients that are dissimilar in their characteristics in their sort of fuzzy matching.

So patient number three does match patient number four with dyspnea, but hypoxemia, angina pectoris, bradycardia, these are all things that don't really exactly match very well with the patient number four's features. And so we've done this basically for all of the patients in the cohort that we identified as having a U09.9 code as Emily had mentioned before, insights that had specific data quality criteria. And so this is a subset of the N3C cohort that we felt was amenable to this type of calculation. Also, it should be noted that not only does not every site contribute patients coded with U09.9 code, but if we were to try to do this with all of the patients in the entire system, our computational needs would probably exceed our capabilities.

It already takes sometimes several days to run these analyses. So it's basically all by all comparison of all the patients that we include here, U09.9 codes. And what we've been able to do is to do that within individual institutions and then compare those patients also across several other institutions. Next slide please. So this is the result of taking those all by all patient similarities and then clustering them based upon the similarities of the patients within each pairwise comparison. And when we optimize the algorithm, we came up with six clusters and they come into sort of predominant symptom family as shown at the top. So cluster number one is just all our severely afflicted patients that really have phenotypic features in all of the categories.

Whereas cluster number two is really focused on pain and pulmonary features. So you can see for example, there are some lab values that are of interest such as hyperglycemia, but we really see a lot in that sort of second orange row of pain, fatigue, chest pain, but less so in some of the other areas. And then again, also pulmonary features such as hypoxemia and cough. The third cluster is focused on pain and neuropsychiatric conditions and it's sort of a little bit the inverse as you can see there in the yellow goldenrod color where we see we have a lot of features from the neuropsychiatric category with headache, insomnia, depression, abnormality of

movement. And this is a very simplified overview. Each of these categories has hundreds of different symptoms that are represented.

Similarly, we have some of our more ENT and pulmonary milder patient cluster number four. Again, you can see there's some cough, but less so on the hypoxemia and a little bit of pain and neuropsychiatric as well as the GI conditions. And then cluster number five is pain and cardiovascular issues with tachycardia, hypertension, palpitations. And then cluster number six is really quite severe again, but has a lot more pain and abnormality of labs than we might see in some of the other clusters. While these kind of overviews may show some similarities and some differences, when you look at the actual individual features, they do sort of optimize to these six clusters. I'll review those in a second on the next slide please. Thanks. So the takeaways from this approach is that we've been able to take each patient and transform their electronic health record data into these human phenotype ontology terms, perform an all by all comparison of each patient's features, and then cluster these into the most like patient groups, which revealed six categories.

As I mentioned, we have this clusters one and six are most severe, the wide variety of symptoms in the different clusters. But cluster six has much fewer abnormality of the various lab values than cluster number one. A pulmonary cluster, a neuropsychiatric cluster, a cardiovascular cluster, and a pain and constitutional symptoms cluster. These are similar but a little bit different than some of the other approaches that we've heard already in this session. And so it's been a really great way to look at these different approaches and how they compare. The limitations of this approach is that not all of the clinical data maps effectively to the human phenotype ontology. So we are missing certain characteristics that just simply aren't represented in the human phenotype ontology. As I mentioned earlier, not all sites are using U09.9, which was used here as the primary inclusion criteria for doing the comparisons.

And as I mentioned, it's also computationally expensive. I should also mention too that not only does not all clinical data map effectively to the human phenotype ontology, but we also know from prior work and from the literature and from many news reports that patient reported symptoms are not necessarily in the electronic health record. They are however represented in the human phenotype ontology quite extensively. And so one of the main goals next for this work is to actually include the patient survey information into the context of this patient similarity clustering approach so that we can take the patient reported symptoms into account, which would be quite expressive using the human phenotype ontology. Next slide please.

Okay, so I'm going to go on to talking about a fourth approach that we've taken in the National COVID Cohort Collaborative, which is really quite different than the approaches that we've shown you, that Emily and I have shown you to date, but is similar to the one that we'll hear about just after me in a second. And so this approach uses something called topic modeling, which basically identifies topics of co-occurring conditions. And so here, what you can basically see is, what essentially it does is it treats the coding systems that are attached to each

patient's record, similar to what we did with the human phenotype ontology approach. It tries to understand which of those co-occurring conditions co-occur the most together and are maximally important in that co-occurrence, so essentially creating topics of these co-occurring conditions.

And the way that you read these displays is that at the top, the displays show the overall usage of that topic. So here the topic, we could just name it by its most frequent contributor, shoulder joint pain, but that topic is essentially 30th by overall usage. You can look at the font size as the probability of the top of that condition in the topic. As mentioned here, 50,000 conditions don't actually fit in the word cloud. So there are many less frequent occurring topics that are not represented on this display, but the size of the font is the sort of probability that you would find a patient with that feature in this topic. If it's more blue, it's more specific to the topic. So relevance. So things that are in that lighter color, kind of orangy color or lighter color are less specific to the topic.

So for example, essential hypertension might be found in very many topics, whereas things like the shoulder joint pain or some disorder of musculoskeletal systems, some of the dark blue ones are more specific to this topic. So that's a little bit of background about topic modeling and that will also, as I mentioned, be important for the next talk as well. Next slide. So topics are used to varying degrees by different sites and reveal expected trends. So how to read this diagram is basically all of the different, so the topics are shown as little squares that are either yellow or blue. The different common data models that is the source data coming from each of the individual sites are shown in the green, orange, blue, pink and green at the very bottom with each individual site sort of randomly displayed there on the bottom. So we can look at each site's topics as a column.

The topic usage goes from the most commonly utilized or sort of occurrence-wise being towards the bottom and the lesser. This list is chopped off quite a lot towards the top. You can start to see some patterns here. For example, if we look at this one that's on the bottom T4, we can see that there are, we're looking at things like cloudy urine and normal sinus rhythms. And so this topic is identified across sites and you can see that by its relative usage there. We can also start to see things that are patterned at specific sites. And this is really interesting. So if you look at those ones that are sort of a whole bunch of blues in a column over there on the left, those all come from some of the pediatric sites that happen to have a lot of commonalities in the way in which they represent pediatric conditions.

And so these are expected kinds of features and you can see that in some of the examples in the pediatric ones such as attention deficit hyperactivity disorder, and these sorts of things that you would find in children. And so this is good, this is a good way of checking to make sure that we're identifying topics that we would expect to be commonly utilized across sites and also revealing site-specific patterns that might be specific to the individual data source types such as OMOP, ACT, PCORnet, or OMOP, or TriNetX, but also to the pediatric population versus adult populations that we see coming from different sites. Next slide please. And so this is some of the results that we wanted to highlight. Here, what we've done is we've shown on the left, these are COVID positive patients

compared to control patients, and on the right, they're PASC positive patients compared to controls as defined by the U09.9 code. Or in some cases, we've also looked at long-COVID clinic visit patients.

So what's shown are really sort of the overrepresented top weighted conditions that we see in PASC and COVID patients compared to the controls within six months of infection. And these are the estimated odds ratios on the x-axis. So what you really want to look at is what do you see in the upper right corner of each of these, which is basically showing the most significant and the highest odds ratio of seeing these conditions as being overrepresented in PASC or COVID relative to controls. And so you see things that you would expect. So for example, we have viral pneumonia as an increased odds ratio with a high P-value, and that's to be expected. We see viral pneumonia in COVID and post COVID patients. What we want to look at though is over there on the right, some of the features that we are revealing as top weighted conditions in our PASC patients.

And you also see some things that you might expect. So we have chronic cough, we have postviral disorder, but we also see some things that we would see, we would expect from patient reports such as sensory disorder of smell or taste and chronic fatigue syndrome. And what's interesting are some of the features that are being revealed such as findings relating to attentiveness, malaise and some of the neurological features as well as some of the lingering things such as hypoxemia and abnormal breathing and other pulmonary effects. These really reflect some of the same kinds of long-term features that we've seen in the other clustering methods as well. And so this is sort of validating of the approach.

One of the reasons that we wanted to use this approach is because we wanted to get after the temporal trends notion that Emily had suggested earlier. I'm going to show that on the next slide. So just very briefly, and this is very early days. I know that we're running low on time, so I'll just very quickly state that we can basically additionally attribute segments of a patient history to that individual topic usage so we can understand how do patients move from one topic to another. So next slide please. And so this is the result of that very preliminary work just to try to understand. Out of 5,400 sex, life-stage, and epic wave-specific contrasts that we tested in these topics, 265 were significant. This represents 65 distinct topics.

So essentially we've now discovered 65 topics that are significant in terms of patients moving from pre-COVID diagnosis to the post-acute diagnosis phase and into the multiple phases of long-COVID post-acute phase. And so what we're really trying to understand is how do those symptom features change over time using this topic modeling approach. I think that is the last slide. So just briefly, again, because we're low on time, topic modeling identifies co-occurring terms and associates them with one or more topics. The commonly reported features are identified. We see demographic specific topics such as weakness and musculoskeletal issues in young PASC patients, hair loss in seniors and female COVID patients, and scarring pulmonary conditions in COVID adults and males.

We see pulmonary issues as well as interesting signals in pediatric patients. And interestingly, non-COVID infections are increased in COVID patients as well. One of the caveats of this approach is just the very large number of topics and clusters and how do we relate those to other approaches. Some of the other issues are that some of the significant clusters relate to non-COVID-related care that may be disrupted such as well child or normal pregnancy. Topic usage varies by site, so the results may not apply equally to all sites. And with that, I'll hand it off to our next speaker. Thank you.

Fei Wang:

Thank you. Let's go back to the PCORnet exercise. We have also done the subphenotype work as Emily and Melissa introduced, and the method we use is topic modeling. I think Melissa has introduced a lot of background about that. Our paper was published on Nature Medicine December last year. Next slide. So this is the pipeline of our approach. So within the PCORnet network, we leveraged the two clinical research networks. One is INSIGHT, as Tom introduced, that covers patient population in New York City area. We also pick another very different clinical research network that is OneFlorida+, including the patients from Florida, Georgia, and Alabama. We pick these two geographically and demographically a very different population as a way of doing validation so that we want to make sure the subphenotype or clusters of long-COVID are replicable.

So the method includes four steps. The first step is we actually look at all the lab tests confirmed, the COVID positive patients. We look at their 30 to 180 days after COVID confirmation to see if there is any new diagnosis incident or in that time period. So by new incident, it means that we don't have any prior records showing that the patient have this diagnosis. It only appears in the 30 to 180 days after COVID positive confirmation. And then we look at the co-occurrence patterns of these new incidental diagnosis because the potential clinical heterogeneity you have seen. So we are expecting a pretty massive co-occurring pattern. So we use topic modeling to first learn a frequent co-occurring patterns of these newly incidental diagnosis as topics.

And then in step three, we learn the patient clusters or the long-COVID subphenotypes on those topic induced patient representations because that can give us a less noisy and more normalized and reliable representation space. And then in step four, we do clustering on top of that to identify the subphenotypes. Next slide. This is the inclusion-exclusion cascade, like I said. We leveraged lab confirmed COVID positive patients records and we do investigation on adult population and we exclude patients without any records in the baseline or the follow-up period. We also exclude patients who don't have any new incident or diagnosis in the COVID period. I mean, long-COVID period and so on and so forth. We do the same inclusion-exclusion cascade on both INSIGHT and OneFlorida database. Next slide.

So this shows the four subphenotypes we identified. I'm sorry, the left figure showing the topics is flipped. But through some numerical optimization, we identified 10 dominant topics. And these topics covers different co-

occurring diagnosis according to the disease area. Some of them focusing on cardiovascular conditions. Some of them focusing on respiratory conditions and neurological conditions and so on, so forth. And then we derive the patient representations on this topic space as a 10-dimensional representation and we run heart clustering on top of that. The right figure shows the four subphenotypes we identified from the INSIGHT network that's New York City area where we see the four subphenotypes.

Number one is dominated by cardiac and renal conditions. Number two is dominated by respiratory, sleep and anxiety conditions. Number three is dominated by musculoskeletal and the neurological conditions. Number four is dominated by digestive and respiratory conditions. The subphenotype one and two are more prevalent than number three and number four. I saw there is a question talking about we should first separate the disease versus the syndromes where the disease has more clear diagnostic criteria or ideologies while syndromes are not. But actually interestingly from our analysis we found across those four subphenotypes, so subphenotype one includes more conditions following the disease type of condition where a lot of the conditions you have clear diagnostic criteria and the etiologies.

But for the other three subphenotypes, they are more like these mood disorder, these kind of syndrome or symptom related disorders which are more challenging to manage, especially as long-COVID. So we kind of tease them out as three main patterns. This is from the INSIGHT Clinical Research Network. Next slide. We replicate that same process on OneFlorida network. Again, the left figure shows the identified topics. This figure is in the right direction. But I mean, in short, I mean these topics are highly overlapping with the topics we identified on the INSIGHT network, which means that these co-occurrence patterns are kinds of similar, although it is from two very different clinical research networks. Same observations for the four subphenotype, which you see the cardiac and renal, and respiratory, sleep and anxiety as two dominant subphenotypes on OneFlorida.

But as you can see, the prevalence of subphenotype one, the 25% is much less than the first. I mean, the subphenotype one prevalence in INSIGHT which is 33%. One of the reason when we dig out is that we find a lot of the patients that falls into subphenotype one. I mean, doesn't matter in INSIGHT or OneFlorida+, they've got their COVID confirmed during the first wave where New York City is the epicenter and we know that lots of the patients who were infected in the first wave, they are associated with more severe clinical outcomes in the acute infection period, which in turn leads to some more severe long-COVID conditions or diagnosis. So that explains the difference because OneFlorida is not the center for the first wave, but really for the second wave. So you see a lot of large prevalence on subphenotype one where a lot of the patients got their COVID confirmed, I mean, after the first wave. Next slide.

So the takeaway of this work is that this PASC or long-COVID study is really clinically heterogeneous and diverse. The waves and also the geographical difference can make the composition of the different subphenotype of patterns differently. But overall, the patterns can be reproduced, I mean in different area. And for identified

subphenotypes, they have a distinct demographic characteristics and also different severity in acute infection period. They are associated with preexisting comorbidities. I want to emphasize this relationship with a preexisting comorbidities are just associations and there is no causality implications. But we still hope this subphenotype work can be way to help us deriving hypothesis on studying the mechanism of long-COVID and further can inform some of the treatment development. So that's all.

Christine Bevc:

Great. Thank you Dr. Wang. All right. Now we're going to bring back Dr. Carton to provide a brief synopsis, just a summary to recap where we are before we move into our quick set of discussion questions with Dr. Megan Fitzgerald. All right, Dr. Carton.

Thomas Carton:

Yeah, thanks Christine. I'm just going to summarize very, very briefly so that we can get into the Q&A with Megan and then continue the Q&A with the audience. Thanks everyone for the questions. We've all been trying to answer the questions rapidly as they come in, and we certainly appreciate all of the feedback. What we just saw in terms of the previous presenters were methods that looked at individual manifestations of PASC, whether they be from the U09.9 work that Emily described or the specific rules-based computable phenotypes to specific diseases that I described. And then we also spent a lot of time looking at the co-occurrence of various organ systems in the manifestation of PASC via either clustering, subphenotyping, topic modeling and so forth. And we looked at similar and also different methods in the ways to do that work.

And the presenters did a nice job of comparing and contrasting the different methods as they presented. In the results, we saw similar outcomes related to organ systems, cardio, whether they be cardiovascular, pulmonological, neuropsych and so forth. So thanks to each of the presenters for both presenting that work and then also working to differentiate and to explain to the audience the differences in those approaches. I'm going to get us started with a question for Megan. Megan, the first question for you is, why do you as a person with long-COVID think this EHR research is important and valuable?

Megan Fitzgerald:

So EHR research can give us a window into the reality of the healthcare experience for people with long-COVID. The EHR database isn't a carefully controlled trial with certain inclusion and exclusion criteria. We're kind of all in there. So EHR can really reflect the real-world experience of people with long-COVID who seek healthcare. And also because there is such a large database in EHR, you can see patterns of the data that wouldn't be apparent in smaller studies. So if there's an issue that's affecting five to 10% of people with long-COVID, I mean, that's a

pretty significant amount of people. But if you have a study with only a hundred people, that might not reach any level of significance in a smaller study, but you might be able to see that in the EHR. And so the EHR can be used to generate testable hypotheses in a lab or a clinical setting because of just the huge statistical power that it has because the sample size is so big.

Thomas Carton:

Thanks Megan. Any of the presenters want to offer any thoughts or comments to this question or answer?

Melissa Haendel:

Sure, I can go ahead. I mean, I think as I mentioned earlier that, and we've really appreciated the inclusion of patients in our actual day-to-day work, it's been a really wonderful partnership. I do think that EHR data has its specific biases as all data sources do. And in the context of EHR data, it's just but a window into the clinical interaction, as Megan had mentioned. And so one of the things that we really want to understand is even though we have a lot of data and having a lot of data can be helpful, it's also very messy data and it's very biased towards that specific clinical encounter context. In the context of the COVID pandemic, patients don't always seek care that they would've normally because it feels more dangerous to go and get care for regular things like pregnancy or well child checks as I mentioned.

At the same time, we also see the most severe patients coming to receive care because they necessarily have to. And so we really, really need to think more about putting the patient back together again as I like to call it, where we seek patient reported outcomes, wearable data, imaging data, and other kinds of data that really help us understand the characteristics of the patient's health in their life, not just through the lens of their clinical encounters. And so it's with these caveats that I think we have to just ethically and data science-wise need to really think about how to best use these data and that context of that bigger picture of the patient's care and health trajectory.

Thomas Carton:

Thanks, Melissa. We got a good number of those questions in the Q&A and the chat as well. In terms of length of time under surveillance, identification of patients for follow up. We answered a good number of those questions in terms of ways that we do our best to mitigate those biases that are inherent in these data that are collected for healthcare purposes and used secondarily for research. We put those out there in each of our presentations and dissemination. One of the things that I think is useful in this understanding is that those inherent biases exist across each of the different sites that are participating and across each of the different patients that

contribute to the data. And one of the values of the EHR data that Megan had mentioned is the size, the scale, the speed and the ability to evaluate trends over time on large samples.

Some of the things that we've learned in our collaborations with both the PCORnet groups and the N3C team as we respond to queries that are brought to us exogenously. Like today, we presented a lot of the work that was developed by our clinicians and data scientists and patient teams, but we also answer questions that come from the community and we present these answers as individual cohorts and then compare the results across. And oftentimes, despite the biases that we've spoken to, we've seen very similar trends across our query responses across groups. Now Megan, I have another question for you. The panelists today shared findings about the different subphenotypes of PASC and the methods that were used based upon these EHR data and studies. From the perspective of the patient community, how can this understanding of the subphenotypes of PASC be helpful? For example, how can they be helpful in getting appropriate healthcare?

Megan Fitzgerald:

I don't think it's a surprise to me as a patient that certain symptoms clustered together that really resonates. And it's going to be interesting to look at this work and follow it as it develops to see if there's a temporality of this clustering over time. But in terms of the value to me as a patient who might be seeking healthcare, I know what my symptoms are, but it can be hard for my doctor to know or any doctors I go to to know what tests to order. So looking at the comorbidities in particular that are clustering with these different groups of symptoms is really valuable in terms of informing the clinical community. These are the tests that should be ordered in these patients so that patients can get the right care and that things that might be treatable, like aspects of their condition that might be treatable, can be treated. Because long-COVID, I don't know, COVID can break our bodies in so many different ways.

So even if you have hyperlipidemia, maybe that's something that can be treated if that correct test is ordered. So I think that that's an important aspect of the clustering in terms of patient care. But I think it's also really important and has implications for clinical trial design and interpretation of clinical trial data. Because it might be the case that long-COVID patients have a respiratory symptom cluster, might respond differently to the same treatment than those who cluster principally with gastrointestinal symptoms. So I'm hoping that this work will have an impact on clinical trial design because we don't want to throw out the baby with the bathwater if a treatment does work for a certain symptom cluster of long-COVID, even if it doesn't work for everybody with long-COVID.

Thomas Carton:

Thanks, Megan. I'm going to ask you one more question and then you have some questions for the panelists as well. So this question is a bit more open-ended, but of really significant interest. What questions do you think the patient community may have about the findings shared today?

Megan Fitzgerald:

So Melissa started talking about some caveats of the EHR data. I'd love to hear more from the other panelists about caveats of working with EHR data because I think a lot of us as care seekers, we haven't always had the most ideal experience with the healthcare system. I wonder how that can be reflected in the EHR data.

Emily Pfaff:

Yeah, I can take a crack at that. There are a lot of caveats and probably could be here all day talking about the caveats. But I think some of the questions that came up in the chat really reflected one of the major ones for long-COVID specifically, which is that many of the symptoms that many long-COVID patients talk about either don't have diagnosis codes, or if they do have diagnosis codes, those diagnosis codes may not be used as frequently as they should be. Diagnosis codes are sort of an enemy and a friend at the same time and that they're really useful for analysis, but they're really designed at least in the United States for billing purposes. And so because of that, they have never been intended to be a complete case history of what patients come and report their symptoms to be, and so we can't assume that they're complete accounting.

For that reason, some of the stuff that's come up in the chat like PoTS and MCAS and even chronic fatigue are not necessarily reflected in those codes. And because they're not reflected in the codes, they may not end up in our analysis. So that's something that we have to keep in mind, that when we report out the results of these analysis, we know that they are not complete in many ways because of those gaps. At the same time, I think that there is utility in what is available there. And as long as we're not using EHR as the only way to investigate long-COVID, I think it can be additive to the many, many other streams of research in this space.

Thomas Carton:

Thanks Emily. Another thing that I'll add here to that question, Megan also came up in the chat. It was the question about the use of diagnostic codes for COVID positivity and how they can be used to define the case definition for COVID positive patients. And the challenge that comes with that relates to Emily's answer is that it's hard to accurately discern a date with only a diagnostic code because the clinician could be taking history of COVID. And so COVID could be coded as a diagnostic code on a particular day, but doesn't necessarily align with the COVID positivity of that patient. And when you're looking at long-COVID, which has a specific time period that's associated with an index state, a lot of our work is in the 30 to 180 days past COVID positivity.

You want to make sure that you're identifying the date of the infection with the actual data in the record, which is why lab tests are easier to work with in that regard. But challenges were brought up with using just lab tests, especially in this era of home testing. So these are challenges that we as a group of three cohorts are regularly discussing and refining and the definition of even just COVID positivity has changed over time and likely will continue to.

Megan Fitzgerald:

Thanks Tom. That kind of leads into what I wanted to ask next, which was about COVID testing. One of the big problems in the pandemic that affected a lot of people, including myself, was that it wasn't always possible to access COVID testing. I got sick in March of 2020 and just couldn't get tested. That wasn't only the case in the beginning. Where I lived in Philadelphia, it was difficult to access tests at multiple peaks in the pandemic. There was some test scarcity. Are there concerns about healthcare equity issues of this nature in the interpretation of EHR and how can these be addressed?

Emily Pfaff:

So I think there are concerns, and that's something that we've been struggling with is how do we find patients that qualify for our various analyses knowing that we need something to grab onto in the data. But also knowing exactly what you said, which is that not everybody who needed a test or wanted a test was able to get one in the beginning of the pandemic and not everyone who had COVID had access to healthcare or chose to access healthcare at the time that they had acute COVID, and therefore their COVID is undocumented essentially. I think we can deal with this in a couple of ways, in that some analyses don't necessarily require us to have a firm index date for acute COVID in order to analyze long-COVID. So for our U09.9 analysis that I was sharing, we did not actually concern ourselves with the acute COVID phase, we just looked for that U09.9.

I don't recall the exact numbers, so don't quote me on this, but over 40% of the patients with a U09.9 in our data did not have record of acute COVID infection in our data. That's astronomical. So if we had required that test in the beginning, we would've excluded over 40% of the patients with long-COVID. So we're very glad that we didn't in that case. Now that's not true of every analysis. There are some analyses that do really need an index date to hang onto, particularly if we're looking at things like, if we're trying to answer questions such as what is the average length of time between an acute COVID infection and then a diagnosis of long-COVID. Clearly, you need something to go on in the beginning there. So I think that the problem that you're bringing up is an important one and is one we need to keep in mind. In the absence of affirmed solution, I think we need to accept that we shouldn't demand record of acute COVID if it is not absolutely necessary for the analysis at hand.

Melissa Haendel:

I might add that we might also not require a diagnosis of U09.9 PASC either, right? So that's knowing, as Emily had suggested, that many sites still haven't even really adopted the code. So we have to have lots of mechanisms to impute candidate patients that meet certain criteria even when the coding systems are not there for them.

Thomas Carton:

Yeah, thanks. One more point that I'll make and then I got a note that we're going to turn this back to Christine. I'll take us back to the first slide that we all presented when we looked at the full map of RECOVER and that the EHR cohorts are one piece that intersects well also with clinical cohorts that do recruit patients and collect patient reported information. So while we're reporting today on EHR only data, we do interact regularly with clinical cohorts. And as our abilities to do linkages, for example, between the clinical cohorts and the EHR cohorts are enhanced, as RECOVER continues, we'll be able to get answers that really combine those two data systems in the broader arc of RECOVER which will help us also understand even the biases that are within EHR and how we can each cover each other's gaps and learn collectively. So Christine, I'll turn it back over to you.

Christine Bevc:

All right, thanks Dr. Carton. All right, so we have a number of questions in our Q&A, thanks to our panelists for answering over 20 of them already. So definitely check to see if your question has been answered. We have time for one, maybe two questions today. This one goes out to the entire panel. It's a bit of a continuation of the previous question that we were just discussing, and it asks, how do you justify or standardize removing people with likely long-COVID from the controls? So the difference between the controls versus those that you're studying, if you can speak a little bit more to the control group, that would be appreciated.

Emily Pfaff:

Well, one thing that I can say while we were talking about EHR caveats is that the control group can be pretty complicated because what we work with in EHR is the absence of information. So if my EHR, my personal one does not say I have diabetes, we have to assume that I don't have diabetes, but that may not be the case. I may have diabetes and it may just not be recorded in my EHR. So in my example, I might end up in a control group, and that might be inappropriate. For the purposes of most research, we do have to make that assumption that the absence of information means that the patient does not have the condition. However, that's really potentially problematic with long-COVID in particular, because as Melissa was saying, that U09.9 code or other ways of getting it potential long-COVID in EHR is not as well-used as diabetes, which has been around for a long time, is well coded generally in EHRs.

And so it can be even more dangerous to assume that a patient that doesn't have long-COVID in their records does not in fact have long-COVID. So the only thing that I can say as far as that goes is that the different methods that all of the presenters talked about today for trying to identify patients, if we take the union set of all of the patients that are identified by those methods and try to use as many different techniques as possible for identifying potential long-COVID patients, we will do the best job in making sure that patients who have the potential to have long-COVID or likely have long-COVID don't end up in a control group accidentally. Are we going to get every single person? Are we going to get it right a hundred percent of the time? No, we're not. But if we can do our best and use different techniques to identify patients, then we will do better than, for example, just using U09.9.

Christine Bevc:

Would any of our other panelists like to add to that in terms of their data sets and the use of controls?

Megan Fitzgerald:

Can I just ask if there's a way that we as patients can know if we have a U09.9 diagnostic code in our charts?

Melissa Haendel:

We, in the N3C, because of the regulatory approvals are not allowed to re-identify any patients, so we cannot do that. However, in the context of RECOVER, there is an effort to create that data linkage. So it's possible that through RECOVER, through future processes that are currently being finalized, that we may be able to provide not only information about what's in your record, but also how you might fit into some of these classification structures or other kinds of outcomes. It would be great to hear from the patient community about what would be useful but not overwhelming.

Emily Pfaff:

And also in preparation for this meeting, I logged into my own patient portal, which many of you may have access to, like a MyChart or the equivalent. I just went to look to see if I could see what codes were on my record from various healthcare encounters. And it's not, at least the way that I was able to see it, I wasn't able to see U09.9 or whatever the code was, but I was able to see what codes, at least in English language translation that my physician had applied at each one of my visits. So I might suggest that patients who are interested, log into your patient portal and see what you see there in your after visits summaries or on your problem list, and you may find a PASC description if that code is on your record.

Christine Bevc:

With that, we actually have reached our time for today's session. So please join me in thanking our investigators and representatives for sharing this exciting work and joining us today. The FAQ for this webinar is going to be posted along with the recording of the webinar on recovercovid.org. It will include all the answers to the questions that were relevant to today's webinar, so those 20 plus answers that were provided along with those that were submitted in advance. Questions about other scientific topics will be addressed in future webinars, and answers to broader questions about RECOVER will be available in the general FAQ on recovercovid.org. As we close out, we invite you to come back and join us for future R3 webinars as we dive deeper into some of the broad topics discussed today.

Our next webinar will be April 11th. When we return for session three in our series on the Mechanistic Pathways of PASC, this webinar is going to take a closer look at organ damage and reprogramming of host tissues and organs. If there's additional topics that you'd like to learn more about, be sure to jot those into the survey that's going to pop up on your screen. I will wrap up by saying thank you again to our presenters and our audience of over 200 plus that have joined us today. This concludes today's R3 webinar. Thank you.