Transcript

Dr. Christine Bevc

Hello and welcome to our attendees and panelists. I'm Christine Bevc joining you from RECOVER at the Administrative Coordinating Center, and I'll be your moderator for today's webinar. I'd like to welcome everyone to today's RECOVER research Review or (R3 webinar. The overarching goal of the (R3 series is to catalyze the formation of a scientific stakeholder community within and beyond the RECOVER Consortium to foster a shared understanding of the state and science. And provide an educational resource for both RECOVER investigators and the broader scientific community of clinicians, patients, and public stakeholders. After we hear from our presenters, we'll use the remainder of our time to answer as many questions as possible and provide responses in real time within the Q and A. If you submitted questions in advance, we appreciate your submissions, and we'll do our best to address them. As a reminder, we will not be answering questions about clinical care and as Shane's mentioned also, the recording of today's webinar along with an FAQ document based on the Q and A will be posted on recovercovid.org.

Today's webinar continues our ongoing series of RECOVER in action. In today's session, we'll focus on disparities and environmental risk factors in PASC with EHR Insights. Our presenters will address what we know about racial and ethnic disparities as well as community level environmental factors in the post-acute phase among adults. As RECOVER continues to grow, we want to remind our audience that the information presented in this webinar is intended to stimulate collaborative dialogue amongst the RECOVER scientific community as well as study participants and other interested parties. As a disclaimer, the information may be recently published or about to publish, and as such, potentially subject to change. In addition, none of this information should be interpreted as medical advice. Now that we've said that, if you haven't already, we invite you to sign up on the website to receive future announcements and updates on this webinar series.

Please join me in welcoming our panel of RECOVER investigators today. It's my pleasure to introduce Dr. Yongkang Zhang, Dr. Thomas Carton, and our discussant, Ms. Gelise Thomas Esquire. Our first presenter will be Dr. Yongkang Zhang. Dr. Zhang is one of the core investigators in the epidemiology health services research component of the PCORnet EHR Hub Under the RECOVER Initiative. He's also an assistant professor in the Department of Population Health Sciences at Weill Cornell Medical College. Dr. Zhang will set the context about what we know about racial and ethnic disparities in the acute and post-acute phase.

Next we'll hear from Dr. Thomas Carton who will discuss EHR findings as they relate to environmental risk factors, as well as addressing the social factors that they found in the data. Dr. Carton is the principal investigator of the Research Action for Health Network or REACHnet and the executive director of the Greater New Orleans Health Information Exchange. They're joined and rounded out by our discussant, Ms. Gelise Thomas, who will kick off our discussion with the synthesis of the presentations and help move us into our Q and A. Ms. Thomas is an unrelenting advocate for equity and biomedical research and healthcare and the inaugural assistant director for strategic diversity, equity and inclusion and health disparities at the Clinical and Translational Science Collaborative at Case Western Reserve University. She also serves as a member of RECOVERs Community Engagement Committee.

Now onto the presentations. Following the presentations, we're going to open the floor to questions from our audience. And due to the size of today's panel, our webinar will be shortened from our usual 90 minutes to 75 minutes. We're still going to reserve time to answer as many questions as possible related to today's topics and presentations. And with that I'm going to turn things over to Dr. Carton to kick us off. Dr. Carton.

Dr. Thomas Carton:

Yeah, thank you very much Christine and thanks everybody for joining today and as Christine mentioned, I'm going to get us started. I have a couple of slides just to really set the scene for the research that Yongkang and I are going to share. Also make a note that Yongkang and I are presenting on behalf of a very large research team that includes researchers, clinicians, patients, collaborations across the EHR cohorts and with the clinical cohorts. So I want to thank everybody in advance for the contributions here and I think you can consider Yongkang and I as deliveries of this research on behalf of that broader team. Next slide please.

Okay, so this sets the scene of the EHR hub studies within the overarching RECOVER organizational diagram. You can see from the top with the executive committee to the steering committee and

then the community engagement panel, various groups, task forces and committees. And then the hub sites are enumerated down here at the bottom, the adult, the pregnancy, the pediatric and the autopsy hub sites. And then there are three EHR hub sites of which the PCORnet Adult group, which Yongkang and I represent is one of three. There is PCORnet Pediatric group and then there is an N three C group. As well we meet regularly, collaborate with each other. They have seen the results of these studies and have helped inform the framing of the discussion. So thanks to the other EHR cohorts as well. And we will refer periodically, mainly in the Q and A to some of the other hub sites and the clinical cohorts and the work that's ongoing across RECOVER. Next slide.

All right, so now we're unique to the PCORnet adult sites and this just speaks a little bit to the data pipeline, the contributing sites, the repository that exists and then some of the use cases. So there are 41 health systems across PCORnet, which is the national Patient-Centered Clinical Research Network that contribute data to this initiative. And then there is a pipeline process of extract, transform, curate, and then remediate any data quality issues that occur before the data finally arrive in the repository, which is where the work that Yongkang and I are going to present is housed. In terms of the uses, there's visualization layer, we have query capacity and are routinely responding to queries from various parts of the RECOVER program. We're able to identify cohorts and we're able to do specific and advanced analytical science, which we're going to show you guys today. Next slide.

So in terms of the work streams for the PCORnet adult site, these are the four bodies of work that we've generally been working across and between. Today we're focusing on the epidemiology health services research side of our team. There are teams that specialize in machine learning and artificial intelligence and that group actually presented at a previous (R3) seminar series. We've got a group that's doing rules-based computable phenotypes that are based on the results of the machine learning models. That are currently under chart review looking at incidents and worsening of existing disease post Covid Ovid 19. And then as I already mentioned, we've got a team that routinely fulfills queries in collaboration with the three other EHR cohorts. So now I'm going to turn it over to Yongkang and we're going to present on two specific studies. Yongkang is going to do the first and he is going to turn it back over to me. Thank you again. And Yongkang, floor is yours.

Dr. Yongkang Zhang:

Thank you so much Tom. Good afternoon everyone and thanks for joining us to discuss this important questions about PASC. Next please. So as Tom mentioned that we are going to present the two studies related to racial ethnic disparity and the environmental risk factors of PASC. The first study I'm going to present will focus on racial ethnic disparities in PASC using cohort locations for New York City. Next please. So first question is like why we connect the research, why we were interested in this topic. I think a big motivation for this research is that we have observed very significant disparities in acute phase of COVID-19 and this evidence has been very consistent and significant throughout the pandemic. We have observed the disparities from beginning of pandemic. There's a lot of studies have reported that patients from racial ethnic and minority groups, they were more likely to be infected, and they were more likely to be hospitalized, and they also had a high mortality because of COVID-19.

So with this evidence we were wondering will the disparities persist in the post-acute phase? In other words, will we observe the racial ethnic disparity in PASC? And we think there are multiple reasons why this would be the case. First, we know racial ethnic minority patients, they receive the delayed treatment for acute COVID-19 and they also have the lower vaccination rates. Some studies have found that vaccinations can be protective, can reduce the risk of PASC. And also there are a lot of social disadvantaged social vulnerabilities among racial ethnic minority patients, such as more access to medical care and all these reasons, plus many other reasons could improve their patients from racial ethnic minority groups and a higher risk for PASC. So we think we need more evidence to understand this important question. Next please.

Next please. So there are a couple of studies have explored and provided some early evidence on this important topic. So I highlighted the three studies here. The first study used about a thousand of patients from Yoo et al health system based on survey data and they found Medicaid patients actually were less likely to help PASC and there's no association between risk, ethnicity and SDoH to determine health in the PASC. The second study, they leveraged the very big comprehensive claims database of a Medicare Advantage plan and they found that Black patients, Black COVID-19 patients, they were more likely to develop certain PASC conditions such as respiratory failure and cardiac disorders compared to white patients. And the third study I highlighted here, they used the data from VA, veteran data and they found Black veterans this COVID-19, they had higher burden of

certain PASC conditions, such as chest pain and shortness of breadth compared to white patients. And they use ICD 10 codes to identify PASC conditions symptoms in the data. Next please.

So we think this studies provided very important early evidence on racial ethnic disparity in PASC, but there are some gaps limitations of these studies. The first as you can see two of them focused on the very specific population group, Medicare Advantage patients and veterans. And the first study, although they focused on a general adult cohort, whether they use the patient from single health system with a very small sample size. So we think we need more evidence on general adults using a large more comprehensive dataset to analyze this question. And also as you may know, people have used different definitions for PASC. So some based on survey data, some based on secondary data. Some examine more, some examine less. So we think we should be a bit more inclusive when we understand the PASC conditions. We should examine more comprehensive set of PASC condition symptoms to understand the disparity. So that's the big motivation why we conducted research on [inaudible 00:13:00] disparities in PASC. Next please.

So the objective of this study is to examine racial ethnic disparities in the incidence of the post-acute sequelae of SARS coinfection PASC among COVID-19 patients. So by instance here we mean we focused on new conditions and symptoms that were developed after a patient tested positive for COVID-19. Next please. So in this study... So this, as Tom mentioned now we have data from more than 40 health systems, but this study was one of the very first studies we conducted. By that time we only had the EHR data from people in the site in New York City, which called the INSIGHT Clinical Research Network. This network provided us with EHR data from five health systems in New York City. So with this data stack we were able to identify patients who were tested positive for COVID-19 or had a COVID-19 diagnosis between March 1st, 2020 to October 31st 2022. So we included about 34,000 patients with COVID-19. Next please.

So the first key question we need to address to understand the disparities in PASC, what are the PASC conditions and the symptoms? As I discussed earlier, the definition of PASC varies a lot in the literature. So here we used a data-driven method to define PASC conditions. So we started a very comprehensive list of conditions, symptoms that potentially are PASC conditions that potentially can be related to COVID-19. So as you can see we included over 6,000 diagnosis categorized them into 137 groups. And then we use a lot of machine learning master as Tom had mentioned, we have machine learning, artificial intelligence team, we also have dedicated coordination team. So we did a lot of machine learning based data driven analysis. We also considered the inputs from clinicians who have been treating patients with COVID-19 and potentially PASC. We also did some ophthalmology hazard research studies to understand the prevalence incidents of these conditions in our patient cohort. So combining all these methods and analysis, we selected about 44 categories and we also did further refinement based on clinicians increased literature. So our final list of PASC condition symptoms included 25 conditions. Next please.

So as you can see, this speaker presents all the 25 conditions symptoms that we use to define PASC. As you can see, these conditions symptoms impact the multiple organ systems of the body and also as I mentioned, this is a data-driven analysis based on our data. But this list of PASC is largely consistent with PASC conditions used by other studies, other researcher groups in the nation or in other countries. Next, please. So to understand the disparities in PASC, our key variable of interest is patient risk ethnicity. In this study we focused on patients from racial ethnic groups. We included the patients who are non-Hispanic white, non-Hispanic Black and Hispanic. So we were not able to include the patients of other racial ethnic purely because of data limitations because those patients accounted for a very small sample size in our data. But with more comprehensive data from over 40 health systems, now we'll be able to examine disparities in other patient groups. Next please.

So to understand the racial ethnic disparities in PASC, as I discussed earlier, we have 25, 24 PASC conditions. For each condition we did a comparison between non-Hispanic Black and Hispanic to non-Hispanic white group using logistic regressions. We also with the data we have because we not only have data applications, what happened to them after COVID-19, we still had the data on what happened before they tested positive for COVID-19. So we were able to adjust for a very comprehensive to list of comorbidities, covariates, potential cofounders including demographics when they were tested positive for COVID-19 and as baseline comorbidities, because we did a lot of the comparisons, we corrected for the multiple comparison. And also as you may understand that the patient who are hospitalized COVID-19 are very different from those who are not hospitalized.

The patient who are hospitalized, they had more severe conditions. So we did stratified analysis by the hospitalization status following the COVID-19 positive test or diagnosis. Next.

So this figure shows us the main results. As you can see we have two panels here. The left panel presents the comparison among patients who were hospitalized for COVID-19. And the right panel presents results of patients who are not hospitalized, who only received the care in ambulatory or ED settings. So as you can see we have the list of the COVID-19 condition and symptoms grouped by the organ system. And for each condition we had two comparisons. The blue symbol indicates the comparison between Black and white. And the red symbols represent comparisons between Hispanic and white. If the symbol is filled with color, that indicates that the difference is statistically significant after we adjusted for the multiple comparison. So as you can see, this results indicates that there are a lot of disparities, a lot of significant difference in the incidents in the risk likelihood of developing this PASC symptoms conditions across racial ethnic groups.

For example, for headache we would see among hospitalized patients, both Black and Hispanic patients were more likely to develop headache compared to white patients. And chest pain is not an example. Both Black and Hispanic patients were more likely to develop chest pain compared to white patients. Also include diabetes and joint pain. For other conditions, for example, for the abnormal pain, well we only observed the significant differences between Hispanic and white. Again for the non hospitalized group we also observe low disparities. Some of those conditions with disparities are consistent with hospitalized groups such as diabetes and chest pain, but there are some variations heterogeneity between hospitalized and non hospitalized groups. Next please.

So the previous slide presented the disparity analysis for each individual PASC condition. We also did a separate analysis by grouping PASC conditions based on the organ system. So as you can see, these results are pretty consistent with what we found for the each individual PASC group, we observed a lot of disparities between Black and white or Hispanic and white in many PASC conditions and symptoms, in both hospitalized and non hospitalized groups. Next.

To summary, in this study we used a large and generalized vocation sample and we found some evidence that suggests disparities in PASC conditions by patient risk, ethnicity, and the key question then why this disparity could exist. So that's something we want to answer next phase, but some potential reasons as I mentioned, it could be due to vaccination rate difference or could be due to patient received delayed treatment for COVID-19. So we are working on getting robust vaccination data and some other data to understand if this could extend disparities in PASC conditions by risk ethnicity. Next. So there's some limitations of this study I want to highlight.

First we identify the PASC conditions based on ICD two diagnosis code in the EHR data, which we think ICD code can do a much better job to identify conditions, medical conditions, but it may not do a good job to capture the symptoms, such as headache. These kind of symptoms may not well captured in the chart data based ICD codes and also we did our best to understand what happened to patients before the COVID-19.

So to identify the incident, new conditions symptoms developed after COVID-19. But as you know, EHR data from this health systems, it is possible some conditions we identified as PASC patients may have this before COVID-19 infection. And also as you know this is a New York City based analysis. Most our patients live in the urban area, although we have some patients live in Jersey or some other suburban areas, but largely this is a urban area based study. So results may [inaudible 00:22:49] to patients from other areas. I think that's all I have here. I think the next... Double check this the last one I have. Yes, this is last one I have and I will turn this over to Tom.

Dr. Thomas Carton:

Thank you very much, Yongkang. And I'm going to describe the second study which are environmental risk factors as the hypothesized outcome variables are similar to what Yongkang described and I'll reference in some cases back to some of the slides that Yongkang just presented. Next slide please. So this slide courtesy of Weaver et al presents a conceptual framework for the relationship between the environmental exposome on COVID-19 instance and mortality and really divides the concept into environmental conditions by climate and built environment, air pollution and environmental chemicals and intoxicants and mechanisms. Which is viral stability, mixing of populations, comorbidities and immune function. And then carries down to the COVID-19 outcomes,

both probability of infection and severity of infection. So this is the conceptual model that really has been guiding the work and we'll work through these systematically as we move on. Next slide.

Similar to Yongkang, I'll describe a bit about the previous work that's been done and found associations between environmental risk factors and risk of incidence and mortality. One looked at historical particulate matter 2.5, which is an air quality measure and found positive association with higher county level COVID mortality rates. Another looked at different air toxicants, nitrogen dioxide and others and food environment and found that they're associated with increased mortality. Both are ecological studies looking at the exposure at the area level and the outcomes looking at the area level, no study has identified environmental risk factors for PASC at the individual level. Next slide.

This slide outlines two essential innovations of the study. One is the comprehensive data that is required and layering of environmental data with patient level data, which we're able to do in the EHR cohort. The underlying clinical data is at the patient level, individual level with patients nested within their environments and we're able to layer on the environmental exposure measures from secondary data sources. Next slide. So we have two objectives of this study. One is to identify the environmental risk factors that are associated with instant PASC conditions, and the conditions are divided into organ systems and then specific conditions similar to what Yongkang described. And then also to understand the heterogeneity of the environmental risk factors. This study includes data from New York City as Yongkang had mentioned, but also from the OneFlorida Clinical Research Network, which includes a handful of health systems across Florida in both urban, suburban and rural environments. Next slide.

This slide outlines the analytic pipeline that we used to get from linkage to findings and we'll describe each of these pieces in more detail. The first piece was the linkage of the environmental risk factor data with the PCORnet clinical data. Then there was an initial exploration that identified through bivariate analysis by each individual outcome signals within those exposome data sets that we wanted to include in a controlled analysis multivariate regression, again by individual outcome. And then finally reporting the findings of their heterogeneity across and then the associations between PASC and 20 in environmental conditions. Next slide. This outlines a handful of the environmental risk factors that were considered for the analysis, both the natural environment, the built environment, and the social environment. Everything from the air quality to food access and walkability to crime and safety and neighborhood deprivation. Outlines the data sources the years used, which we'll come back to at another slide and other factors, number of variables that are linked in to our RECOVER EHR data. Next slide.

Here we're describing the analytic pipeline. The engineering piece is the linkage. The phase one piece is the univariate analysis where we ran simple regressions for each outcome, and we had outcomes based upon families of conditions. So one of eight which Yongkang described as well as 25 individual PASC conditions as Yongkang described as being derived from the machine learning model. And we used one exposome factor for each piece. So this was a large number of simple regressions and we controlled for demographic and community variables, but the exposures and the outcomes were included one by one to identify signals, which then we brought forward into phase two of the analysis, which again looked at PASC families as well as individual condition groups, but included all of the exposome variables that were included as significant in the simple regression models and again, covariates of demographic and comorbidity variables. Next slide.

This is a visualization of phase one where the PASC families are listed across the top as columns and the various exposome factors are listed as rows. And those rows are identified as groups in the column to the far right food access, and to social capital, and walkability, and other things. And this is where we are identifying specific signals within these univariate regressions that we would then build into the multivariate regression as phase two of the analysis. Next slide.

And this is a slide that's outlining the phase two of the analysis by the PASC families. So this is the eight condition groupings that are listed on the far left of the diagram. They're spelled out by the insight New York City Clinical Research Network and the OneFlorida Clinical Research Network. And then the specific exposomes that reach significant levels are present in the middle column and the confidence intervals are present in the graphic on the far right. And the point that we'll make here and we'll make again in the discussion is really the heterogeneity of associations that we found between the exposome factors, both in terms of the size of the effect, the confidence intervals of the effect, and the varying effects across geographic domain. Next slide.

This slide is designed in a similar way except the column on the far left are not the PASC families or the condition groups, but they're individual conditions. The second column again is looking at the exposome factors and the confidence intervals are on the right. This is not segmented by geography, but again shows vast amount of heterogeneity of the relationships between the exposome and the PASC conditions. Next slide. So in terms of the main findings, most of the exposome factors that were identified as risk factors per PASC were air toxicants, overall neighborhood deprivation, and concentrations of particulate matter 2.5. There were a few built environment characteristics, such as food access that were also significant risk factors. And the findings indicate significant heterogeneity in environmental risk between New York City. Again, more of an urban and suburban environment, and Florida, which is a bit more of a diverse geographic environment. Next slide.

I've also identified multiple direct and indirect pathways that exist. One is long-term exposure to air pollution causing various symptoms and conditions of the central nervous respiratory endocrine or other organ systems. The second is that air pollution can modify individual susceptibility to SARS-CoV-2 infection as well as severity and it can be mediated by several factors here as well. Next slide. The strengths of this analysis are really being able to look at the simultaneous exposure of the various pieces of the exposome as presented in the conceptual model. It was a systematic approach through two phases to identify univariate signals and then multi-variable regression models to investigate and to interrogate multiple exposome factors at the same time. The second is our ability to adjust for detailed patient level characteristics because of the fact that we had individual level patient data that was rich in the capture of both race, ethnicity and comorbidities and geographic residents. And then finally is the comparison of the findings between two large cohorts, one in New York City and one across the state of Florida. Next slide.

In terms of limitations, the first is that the exposome characteristics were grouped at the zip code five level. Essentially at a patient's zip code, which might not be granular enough to estimate some individual risk factors. The second piece, which we touched on briefly when we were talking about the data that was layered in with the exposome data that was linked is that much of this data included exposures prior to SARS-CoV-2 infection and that those toxicants might have changed during the pandemic. Another piece is that we had a lack of residential history data. So the residential history was taken from the patient's clinical record, but taken at one point in time and didn't account for the movement of populations across the geography.

And then a large one which will come up when we discuss the pros and cons of using EHR data for this work is that there were several potential confounders, vaccination status is an important one, that is hard to capture reliably in any EHR data and might explain some of the heterogeneous findings between New York City and Florida. Next slide. All right, finally, just wanted to acknowledge the RECOVER funding source, the MPIs at Weill Cornell that are unable to join us today. Our team of data scientists, the University of Florida and OneFlorida Clinical Research Network. And then our machine learning team that generated a lot of the early evidence that helped us identify the PASC conditions and condition groups that we used for this analysis. So that concludes Yongkang's and my presentation and we are going to pass it over to our discussant, Gelise Thomas to take us into the next phase of the webinar.

Christine Bevc:

All right, and Dr. Carton, I'm going to jump in real quick and just give a reminder to our audience members that you guys have been doing a great job of responding to some of the questions that have been submitted through the Q and A. So please continue to submit those and then after we've gone through with our discussant Ms. Thomas, then we'll return to those. But if you submitted a question, be sure to check back to see if it's been answered. We will touch on a few of those after the discussion. All right, so I'd like to welcome Ms. Thomas to provide us with a brief synthesis of these presentations and help lead off our discussion. Ms. Thomas.

Ms. Gelise Thomas Esquire:

Thank you so much, Christine. These have been fantastic talks. Thank you Dr. Carton and Dr. Zhang. It's my pleasure to give a brief synopsis of both presentations and from my perspective as a research diversity, equity, inclusion, and accessibility professional, I am especially intrigued by the focus on race and ethnicity and socioeconomic status and really just how all of this dovetails into the various frameworks when we look at this from an environmental health perspective. So doctors, Carton and Zhang's work has illustrated the impact of disparities in PASC from a racial and ethnic minority varied socioeconomic status and environmental perspective. Highlighting a few recent studies on disparities in PASC, we learned that there's much more that we need to learn

to better understand the disparities that persist. When we focus on the environmental exposome and COVID-19, the environmental conditions mechanisms and COVID-19 outcomes provide a framework upon which we can potentially contextualize racial and ethnic minority and varied socioeconomic status disparities for PASC.

Dr. Carton set the stage and emphasized the many stakeholders who contribute to this great work by developing analyses for exacerbation of preexisting diseases and investigating racial and ethnic socioeconomic temporal and geographic disparities in PASC. He and Dr. Zhang were able to conduct EHR based cohort studies from the RECOVER program. Dr. Zhang focused on the racial and ethnic disparities in the acute phase of COVID-19 that has been consistent and significant throughout the pandemic. Potential reasons for the disparities were delayed treatment for acute COVID-19, lower vaccination rates, and disadvantaged social conditions. Dr. Zhang identified numerous gaps in literature including Medicare and veterans, general adults, and expanding the set of PASC conditions examined as well as a lack of comprehensive assessment of patients beyond surveys. And he said that the key question that needed to be understood was how PASC was defined and he was able to refine that list to 25 conditions.

For methods again, there was a focus on race and ethnicity, non-Hispanic white, non-Hispanic Black and Hispanic. And the sample size was not more comprehensive since data was not available at this time, each condition had two comparisons, Black and white and Hispanic and white. And some of the conditions were consistent with hospitalized versus non hospitalized groups. Some of the limitations that were highlighted included inconsistent diagnosis codes for PASC and capture of symptoms within the EHR. Dr Carton began his talk with a graphic that illustrated the environmental factors influencing COVID-19 incidents and severity. He talked about previous work done and associations between environmental risk factors and increased mortality of COVID-19. To date, no study has examined environmental risk factors for PASC. Dr. Carton said that comprehensive data is needed so that environmental risk factor analyses can be layered onto that data. The objectives of the study were to identify environmental risk factors associated with development of incident PASC conditions and symptoms, and to understand the heterogeneity of environmental risk factors of PASC across geographic regions.

His analytic pipeline that he used had data linkage, initial exploration findings and controlled analysis. The EHR linked with 200 environmental characteristics was looking at one environmental characteristic to one PASC outcome and he found heterogeneity across NYC metro and Florida as well as PASC associations with 20 environmental risk factors. Natural built and social environmental factors were examined based on zip code. His phase one results included PASC families, so he had eight condition groupings and environmental exposomes. And then phase two results had individual conditions listed as rows with the second column looking at exposomes and confidence intervals. Most risk factors were air toxicants, neighborhood deprivation, a few built environment characteristics like access to food, and significant heterogeneity and environmental risk factors for PASC between NYC metro and Florida. And now that we've briefly summarized each talk and study, let's move to a few discussion questions inspired by this phenomenal research. So what are the pros and cons of EHR and other data sources to answer questions about disparities?

Dr. Yongkang Zhang:

I'll take the question. This is such good question, Gelise. So people let me know from literature, researchers use different data sets to understand the PASC and disparities. Many people like us, we use the EHR data, either EHR from a single institution, single health system, or EHR from multi health systems. Some studies as I described, one of them use the claims data and many researchers, especially in the early phase before this comprehensive, secondary data become available. They leverage the survey data and interview data. For example, if I'm a physician, I will call my patient after the discharge from a hospital because of COVID to understand new conditions symptoms developed after the discharge from hospital. So a lot of data have been used to understand PASC and disparities in PASC. So I think there are pros and cons.

EHR data first, it's timely available compared to claims data, as we know claims that office has a lot of lags because they're from the insurance companies many claim's data. So they not become available in next 12 months, six months for example. And another limitation of claims data, like many claims data don't have lab tests, which is a key part for us to understand the COVID-19 status. Of course you can use diagnosis, but diagnosis is not reliable for COVID-19. We don't know... Diagnosis may be like few days later after a positive test. If a patient never go to hospital after a positive testing in some street by street testing spot by street, you probably will never see a diagnosis.

So we think actual data is timely available with comprehensive lab data, which allows us to understand who have COVID and when. And also a lot of numerous diagnosis procedure to understand not only what happen to patients after the test of positive, but also what happened before they tested positive. [inaudible 00:43:52] very comprehensive longitudinal information of each patient to understand their medical history and their new condition symptoms. But limitation is that EHR data, you only have EHR observation when you go to hospital. If a patient never go to hospital, never go to physician office, we don't know what happened to them. Maybe some patients they have developed a very severe PASC conditions, but because they have transportation issues, because they don't have primary access to healthcare, their information, their record will never be seen by researchers in EHR data. And also another thing is that EHR data only represent information in physician office, primary care, ambulatory care, and acute care. So we really don't have EHR from the post-acute care or long-term care.

This is important for some subgroups such as older adults. As we know, nursing homes got hit hard by COVID-19. And if you use hospital EHR data, you don't know what happened to nursing home patients in nursing home because people just don't have access to that EHR data. So that's pro and cons. Ideally, we would allow it to combine all these source together. EHR data, claims data and the survey data because as I mentioned, some symptoms may be poorly captured in both EHR claims data. Like physician may not refer to ICD two code for headache or some minor symptoms instead of more severe conditions. So ideally we think we should combine this all dataset. And also as Thomas research mentioned EHR claims medical data to understand the social environment, we need to combine link electronic claims with social data, which are from non-medical settings. So I think this is an important question which really tell us the importance of developing a robust dataset with all kinds of information to understand the PASC related to disparities. Thank you.

Dr. Thomas Carton:

Thanks Yongkang. And Gelise first, thanks for the terrific summary of the findings. Really terrific overview of everything and I agree with what Yongkang had mentioned. I'll add a couple of additional points. One is that where EHR data has some limitations in terms of generalizability and external validity for population science, it does allow us to interrogate trends over time. Which are generally consistent even within that smaller population.

The lab test data is a very important point that Yongkang made when comparing EHR data to claims data specifically for PASC, when most of the work that we've done has identified the COVID-19 infection as the index date, that set the time for studying that patient cohort into the future. And a diagnostic code could be a diagnostic code for a history of COVID. And the date that the diagnostic code is coded might not be exactly temporal to the onset of infection, whereas the lab test data does allow us to do that. So that's an important point. And then one additional point that I'll make given that we're talking about disparities and looking at race and ethnicity is that EHR data generally has a better capture of race and ethnicity data than claims data. It allows for interrogation of some of those disparities in ways that other data sets might not.

Ms. Gelise Thomas Esquire:

Thank you so much, Dr. Carton, Dr. Zhang. Let's look at this from a patient's perspective. What could a patient do to help move forward the understanding of PASC as a racial and ethnic minority or someone who has been impacted by environmental risk factors?

Dr. Thomas Carton:

Yeah, I'll take this first and then Yongkang, you can chime in. I think one of the real pleasures in working with the RECOVER program and the multiple layers of RECOVER, and I'll just make a point too that we're presenting data from three EHR cohorts. When we looked at the teams in the slide that outlined the organizational chart, there were 20 or 30 other cohorts that are also doing analysis that's in scientific investigations that are similar to this. And then the connection between the patient advisory groups and the various levels of ways that patients are integrated into the RECOVER program is a real strength both in terms of bringing questions to the researchers, but then also disseminating the findings of the research.

So Gelise, to the question the dissemination of this research is incredibly important to get out to patient populations so that specifically related to the exposome and the risk factors, patients could get a sense for where they sit within the environments in which they live. People are just are well aware if they're living in environments that have poor walkability or limit access to food because they have to deal with those issues every day, in terms

of being able to get access to those food environments. Certainly maybe a little bit less aware of some of the air quality issues, but in some extreme cases certainly are aware of those issues.

As these findings get disseminated and participants and patients and the community at large understands the fact that these exposome or environmental risk factors exist. If they're in one of these risk riskier and environments, that's important for them to be able to know both in terms of the way that they care for themselves and seek care. I think that's one, there are several other potential avenues to answer that question Gelise, but I thought that answering it from the perspective of the connection to patients within RECOVER and dissemination through webinars like this made sense. Yongkang, anything that you'd like to add?

Dr. Yongkang Zhang:

I think Tom, you have done a great job of explaining this. I totally agree with you. It's by disseminating this research and through our collaborations with other RECOVER groups and also in PCORnet, we have very close engagement of patients groups. And by disseminating these research results through these channels, it really tells patients the importance of their environment and how does that relate to their health, especially the PASC condition symptoms. I think that's a key part for patients to take care of themselves and reduce the risk of having PASC after a positive test of COVID-19.

Ms. Gelise Thomas Esquire:

Thank you, Dr. Zhang and Dr. Carton. I think for our final question before we move to the audience Q and A, what are the policy questions and/or implications of these findings about disparities in PASC?

Dr. Yongkang Zhang:

Do you want to start first?

Dr. Thomas Carton:

Yeah, I can start. I think that I'll... From a policy perspective, the ability of this research and other research across the RECOVER program just to highlight these disparities in terms of racial and ethnic disparities and social environmental disparities is of critical importance. And also not just to highlight, but to enumerate it and to quantify it and to be able to publish it and disseminate it. I think that too many times the scientific literature, it needs to speak very clearly to what the factors are and how to quantify those so that policy makers can understand and can react to that. In terms of specific policymaking or other things, I think I'll hesitate there because that's just not my area of expertise. I think where I'm coming from for this Gelise, I consider it our job to get these investigations out into the literature, to promote these that policy makers then who have expertise in how to respond to and deal with these issues can do such. Yongkang.

Dr. Yongkang Zhang:

Yes, thank you Tom. I totally agree with you. I think I'll discuss a little bit more from a clinical care delivery perspective and also the policy perspective. I think there were lots of efforts and policy efforts during the pandemic, either from federal government or from state government or from local government to address disparity indirect. Some policies, not health policy, but from other policy to help patients from minority groups or from low socioeconomic groups to improve the health or improve living status during pandemic. And as we are entering a post pandemic year, I think first it's important for policymakers and providers to realize that these disparities that we observed during the acute phase still exist. So there should be policy from different levels, federal government, state government or local government to address disparity in the post-acute phase of the COVID-19. And as we discussed this PASC actually impacted the different organ system of patients.

So from health system perspective, usually think about if a patient who had the COVID before presented to health system. What kind of resources they would need to treat the patient with PASC, because PASC is not a single condition symptom, it will be multiple organ system, it'll impact multiple systems. So how does that inform the planning and the allocation of health resources, especially to the regions locations that have a much higher prevalence or incidents of COVID-19. Those regions may have a higher proportion of patients this PASC that will present in health systems. And are we ready to treat the patients with PASC given the resources they have? And as we know, COVID-19 really impacted the staffing employment of the healthcare professionals. So do we have a shortage of professionals to treat the patients PASC? I think those are very important PASC policy questions we need to answer to really help patients recover from the COVID-19 infection, especially those from racial ethnic minority groups.

Ms. Gelise Thomas Esquire:

Well, I think we've heard some direct ask and actions that our listeners on the call and those who will listen to this on demand can take up with their health systems or academic medical center. So thank you so much Dr. Carton and Dr. Zhang for your insights. And now I pass the virtual mic back to Christine.

Christine Bevc:

All right, thanks Ms. Thomas. All right, so now we're going to take a few minutes to answer our audience questions for our attendees. Please continue to submit your questions into the Q and A. Our first question was submitted in advance and asks, have you found that long COVID patients who were born or reside in large cities are worse off than ones from small cities?

Dr. Yongkang Zhang:

Tom, I can throw some thoughts and you can add more. So the first study we conducted about racial ethnic disparities, most patients are in New York City area, which is a very urban area as you can imagine. But I think in Tom's study really highlights some things for this rural/urban differences, although we did not compare the PASC differences between rural and urban. But as Tom mentioned, the Florida and New York City, which are very two different geographical regions. New York City, we have very good public transportation, we have pretty good food access. But for Florida, the environment, social environment will be totally different. And as many people have found that social characteristics can be with factors for PASC. So considering these differences in environmental social conditions, we would imagine there could be a differences in PASC conditions symptoms between people from urban areas and the people from rural areas. And I think we will be able to enter this question given the data that we have, which cover patients from different geographic locations in the US. Tom would you like to- [inaudible 00:56:51]

Dr. Thomas Carton:

Yeah, I agree Yongkang. I mean Christine, it's not a question that we directly posed of the data. As I mentioned, we really grouped the environmental risk factors at an area of zip code five and then identified the exposures to the patients within those regions. We didn't explicitly identify urban, para urban, rural as variables of interest for these studies. Although as Yongkang mentioned, we can. I think to the degree of which these environmental risk factors vary from urban to para urban, suburban and rural environments, we would be able to interrogate those. But in terms of the way that we handled it for the study that Yongkang and I described on exposome, we really just nested those variables within the zip code of residents and didn't specifically interrogate by size of city as your questioner has asked.

Christine Bevc:

Thank you. Our next question comes from our audience member and asks, how do you check for inaccuracies in the data? And this was something that Dr. Carton, you'd started to answer and we're continuing that discussion a little bit more now here in the Q and A. So if you could expand on your notes there.

Dr. Thomas Carton:

Yeah, that sounds great Christine. It's a great question. I mean data quality is of incredible and importance. And then the fact that we have EHR data from such a large number of patients and sites really does allow us to do a good amount of data quality and interrogation. I briefly mentioned some of the checks that we have in place and we do data quality checks as the data is moving in to the repository, as I discussed with the earlier slides. And then even within the confines of specific studies, we're able to interrogate issues of plausibility and completeness. Plausibility, are you seeing breast cancer diagnoses among men would be a good example of implausible EHR data. Completeness, do you see extreme changes in a person's record over shorter periods of time? And there are other various data quality checks that we're able to perform.

And then we have the ability to go back to the sites and present the data quality challenges back to the data contributing organizations and ask them to investigate those and to let us know if it's a problem with their source data or if it's something that just was a mistake in the transmission. Now it's harder, and the question was direct about how do we understand just mistakes or inaccuracies in the coding? And for very specific individual level patient mistakes in diagnoses or coding, it is very difficult for us to do and it is a limitation of the EHR data. We're able to interrogate general patterns or quality issues certainly by health system, some cases by site within system, some cases by specialty within system. But the data science that we do and have access to is not good

enough to identify just coding errors at the physician's office. That's one where we can identify general trends, but individual "mistakes" or send back, we just don't have the ability to do that. It's more of a population science trend based analysis.

Dr. Yongkang Zhang:

Yeah, thank you Tom. I'll add a couple of more thoughts. So as Tom mentioned, there are some intrinsic limitations to the EHR data and we have done our best to address these limitations once I want to discuss. First I'll discuss is the accuracy of COVID-19 status. So as you know, patient can get tested anywhere. Not only in the health system but also pharmacy or [inaudible 01:01:09], which means that some patients will observe who are negative, they never had a positive test, all the tests were negative, they might have a positive test somewhere else and we just don't see in the EHR data. But this limitation addressed can mitigate a little bit by using data from file health systems because if this patient with the multiple health systems all the system capturing our data, we'll be able to do a much better job in capturing the correct status of COVID-19 tests compared to a study using EHR from a single health system.

So there's something we can mitigate by using a more comprehensive dataset, but again, if patient that had a positive somewhere else, we just don't know, but we do something to mitigate this bias or inaccurate soon. Another thing probably more relevant to the first study we conducted is accurate of risk ethnicity information in the EHR data. For those people who use the EHR to understand ethnicity disparity, you may know that a higher proportion of patients actually don't report for some reason or don't report or don't capture the risk ethnicity EHR data. So we have a lot of patients the risk is unknown or other or no information. And for those patients, we really just don't know their accurate risk ethnicity information. So it really poses a challenge for our study.

Of course there are some methods we can use to address this limitation. For example, there are lots of national processing based research to impute the missing or unknown risk ethnicity status based on clinical notes, which will definitely require we use unstructured data to capture patient risk ethnicity. So I think the point is that there are a lot of limitations in the EHR data. There's something we can do to meet the bias, but some limitations may still exist. Even this strategies [inaudible 01:03:09] we have used to meet these biases in the EHR chart data.

Christine Bevc:

Thank you both. And I wanted to also pose this to Ms. Thomas because what can patients do to help ensure accuracy in their own medical records?

Ms. Gelise Thomas Esquire:

Definitely using that self-advocacy lens or muscle, you could request a correction in your medical record. You can review your medical record and if something does not align with your understanding, you can report that and request that correction.

Christine Bevc:

All right, thank you. All right. Our next question follows up on the exposome data. And are there similar exposome data found in other diseases, both viral or non-viral in origin such as chronic cardiovascular disease, pulmonary, other diseases?

Dr. Yongkang Zhang:

Tom, I can start. So part of the motivation of this study is that many prior studies have reported environmental or the environmental people called exposome factors are risk factors for respiratory disease. For example, for COPD or some other things. And also some other food environment is also important for patients with diabetes, building environment is also important for heart failure or some other conditions. So there's very robust evidence on the association relationship between exposome or environmental factors and a lot of chronic conditions. There are also very robust evidence on exposome environmental risk factors and acuities of COVID-19, like a rate of positive infection, hospitalization rate, and mortality. A couple of studies have used [inaudible 01:05:09] data as Tom mentioned, to establish association between PM 2.5 and the mortality of COVID-19. So all this give us a motivation to think there could be association between exposome manufacturers and long COVID or PASC. I think that's definitely the part of motivation why we pursue this study in the first place. So I'll stop here Tom, do you want to add something?

Dr. Thomas Carton:

No, I don't really have anything more to add on this one Yongkang.

Christine Bevc:

All right. So our next question asks in the JGIM article, you noted that race ethnicity may be an independent risk factor for PASC, not explainable by racism or socioeconomic factors. Can you clarify what you mean by race as an independent risk factor in this case?

Dr. Yongkang Zhang:

Yeah, that's a good question. First I think the question, it's a very important question and I think also related to another question I saw from the Q and A section. So this is risk ethnicity and the PASC, because there are multiple pathway between risk ethnicity and the PASC. Some of the pathways are direct, there's direct associations between risk ethnicity and the PASC. There are some indirect associations, for example, mediated by some other factors. So the question, when we find the association a significant disparity by risk ethnicity, it could it be just purely because of risk ethnicity or could it be expanded by some other factors?

So as I presented in the study, we adjusted for a set of covariates that could be potentially the confounders. But beyond that, what other factors could also expand racial ethnic disparities? I think we conducted based on some comments we received from reviewers actually of the journal, we conduct additional analysis by adding some neighborhood level social economic factors, like meaningful income in zip area or something else to see if these social economic status could further explain the racial ethnic disparity. And we found that after adding these economic factors, we still observed the very significant disparities by risk ethnicity. That's why we call it could be an independent risk factor for the PASC.

Christine Bevc:

Anything to add Dr. Carton or no?

Dr. Thomas Carton:

No, thanks. Nothing to add.

Christine Bevc:

Okay. All right. Another clarification. So by protein intake, are you referring to spike protein introduction via vaccination status or some other source?

Dr. Yongkang Zhang:

Well, yeah. As I explained already, I think Tom and I are not experts on this question. We do have some environmental health expert in our team. I think by that protein part, I think that primarily means that there are some biological disorders induced by the air toxic environment. I think we have at least one paragraph in our paper to discuss the potential mechanism behind the environmental spectrum PASC. So feel free to check the paper for more information.

Christine Bevc:

Okay, thank you. Anything to add Dr. Carton?

Dr. Thomas Carton:

No, thanks.

Christine Bevc:

All right. Continuing that line of environmental health and environmental risk factors, can you just remind us what was included in your analysis for the environmental factors in the study? Was exposure to mold in there or is that something that's going to be considered for future studies?

Dr. Thomas Carton:

I'll start this one. No, exposure to mold was not included. And we got some other questions in advance of the presentation related to say indoor air pollution, which was also not measured. These were sort of external environmental factors that occur outside of residents that were able to gather from independent secondary data sources. Those types of questions... Also saw in the Q and A, the specific question to mold. That's just a limitation of the data that we are going to be able to interrogate an answer. But I'll bring it back to the clinical cohorts and remind all of us, and then the webinar participants as well that there are other clinical cohort studies that are enrolling participants with consent and asking them direct questions of their behavior, behavioral history environment and others.

Now Christine, I don't know if those studies are asking questions specific to mold. We'd have to dig that up and look that up or talk to some colleagues about that. But there is an opportunity to study those specific questions related to individual exposures that are outside of both the environmental factors and the clinical factors that we can get through the HR that are being investigated through various parts of RECOVER.

Christine Bevc:

All right, great. Thank you. And with that, that is actually going to be our last question. We do have several questions that were submitted. Those are going to appear in the FAQ document for this webinar. That's going to be posted, along with the recording. That's going to appear on recovercovid.org. So please join me again in thanking today's panel for sharing these updates. And thank you to our audience for joining us today. If you have questions about scientific topics not addressed today, may be addressed in future webinars, as well as broader questions about RECOVER. Such as the different cohorts that we have, the observational cohort, as well as our real world data collection. Those are also available in the general FAQ section on recovercovid.org.

So as we close today, I want to make mention that we are going to be taking a short summer break next month and look forward to returning after August with a new wave of findings as RECOVER continues. If you haven't already, please remember to sign up on our website to receive those future announcements and details about upcoming webinars. And lastly, if there's topics that you want to learn more about, please be sure to enter your ideas into that survey that's appearing on your screen somewhere. So thank you to our panelists and also our audience today for joining us. Take a moment to provide your feedback there. And this concludes today's (R3) webinar. Thank you.

Dr. Yongkang Zhang:

Thank you.

Dr. Thomas Carton:

Thanks everyone.

Ms. Gelise Thomas Esquire:

Thank you.