Responses to Participants' Questions

The overarching goal of the R3 Seminars is to catalyze a shared understanding of the research being conducted by the scientific stakeholder community within the RECOVER Consortium. The R3 Seminars and the Q&As typically feature highly scientific material intended for researchers and clinicians. For other audiences interested in these topics, a link to the National Library of Medicine's MedlinePlus medical dictionary is provided at the end of the Q&As as a resource to help in understanding the scientific terminology.

This document provides responses* to questions raised by seminar participants related to the following presentations at the R3 Seminar ***RECOVER in Action: Disparities and Environmental Risk Factors in PASC, EHR Insights*** held on July 11, 2023:

- ***Disparities and Environmental Risk Factors for PASC, EHR Insights***
  Yongkang Zhang, PhD
  Thomas Carton, PhD, MS
- **Discussant: Gelise Thomas, JD, MS**

* Responses may have been edited for clarity.

All Presenters: Questions and Responses

Q. What are the pros and cons of EHR and other data sources to answer questions about disparities?

Responses:

Dr. Zhang: Researchers use different data sets to understand PASC and disparities. We use electronic health record (EHR) data, either from a single institution or from multiple health systems. One of the studies I described used administrative claims data. Other researchers, especially in the early phase before this comprehensive secondary data became available, leveraged survey data and interview data. For example, if I'm a physician, I'll call my patient after discharge from a hospital because of COVID to understand new conditions and/or symptoms developed after discharge. So, a lot of data have been used to understand PASC and disparities in PASC. I think there are pros and cons to using EHR data.

One pro of using EHR data is that the data are timely compared with claims data, which lags because of the time spent on processing claims and making the data available for research. So, the claims data will not become available until the next 6 to 12 months. Another limitation of claims data is these data often don't include lab tests, which are key for us to understand COVID-19 status. Of course, you can use diagnosis, but diagnosis is not reliable for COVID-19. For example, diagnosis may be a few days later after a positive test. If a patient never goes to the hospital after a positive COVID-19 test from a non-hospital testing venue, you probably will never see a diagnosis.

We think EHR data are timely and provide comprehensive lab data, which allows us to understand who has COVID and when they had it. Also, EHR data have numerous diagnosis and procedure codes to understand not only what happened to patients after a positive COVID-19 lab test, but also what happened before they tested positive. That would give us very comprehensive longitudinal information on each patient to understand their medical history and their new conditions and symptoms developed after COVID-19 infection. However, a limitation is that EHR data are only available if a person goes to a hospital. If a patient never goes to a hospital or a physician's office, we don't know what happened to them. Maybe some patients have developed very severe PASC conditions, but because they may have transportation issues or because they don't have primary access to health care, their information will never be seen by researchers in EHR data.

Another downside of EHR data is that they only represent information in the physician's office, primary care, ambulatory care, and acute care. So, we really don't have EHR data for 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 because we just don't have access to their data. So, these are some of the pros and cons.

Ideally, we'd be able to combine EHR data, claims data, and survey data because, as I mentioned, some symptoms may be poorly captured in both EHR and claims data. For example, the physician may not report the ICD-10 code for headache or some minor symptoms. So ideally, we think we should combine all these datasets. Also, as Dr. Carton mentioned in his presentation about using EHR data to understand social and environmental risk factors for PASC, we need to link EHR data with social and environmental data, which are derived from nonmedical settings. I think this is an important question that points to the need to develop a robust dataset with multiple types of information to understand PASC related to disparities.

Dr. Carton: I'll add a couple of additional points. One is that although EHR data have some limitations in terms of generalizability and external validity, these data allow us to interrogate trends over time, which are generally consistent even within that smaller population.

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

Q. 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?

Responses:

Dr. Carton: One of the exciting aspects of working with the RECOVER Initiative is 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 in scientific investigations that are similar to this. And then the connection between the patient advisory groups and the various levels and ways that patients are integrated into RECOVER is a real strength both in terms of bringing questions to the researchers and in disseminating the research findings.

So, to the question, disseminating this research to patient populations is incredibly important, specifically related to the exposome (a concept used to describe environmental exposures that an individual encounters throughout life, and how these exposures impact biology and health) and the risk factors, so that patients can get a sense for where they sit within the environments in which they live. People are usually aware if they're living in environments that have poor walkability or limited access to food because they have to deal with those issues every day. They may be less aware of some of the air quality issues, but in some extreme cases certainly people are aware of those issues as well.

As these findings are disseminated and participants and patients and the community at large understand that these exposome or environmental risk factors exist, if they're in one of these riskier environments, that's important for them to know in terms of the way they care for themselves and seek care. There are several other potential avenues to answer that question, but I thought that answering it from the perspective of the connection to patients within RECOVER and dissemination through webinars like this made sense.

Dr. Zhang: I totally agree with Dr. Carton. It's by disseminating this research and through our collaborations with other RECOVER groups and in PCORnet that we've fostered very close engagement with patient groups. And by disseminating the research results through these channels, it really tells patients the importance of their environment and how it relates 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.

Q. What are the policy questions and/or implications of these findings about disparities in PASC?

Response:

Dr. Carton: From a policy perspective, the ability of this research and other research across the RECOVER Initiative to highlight racial and ethnic disparities and social environmental disparities is of critical importance. Also, not just to highlight, but to enumerate them and to quantify them and to be able to publish and disseminate the findings. The scientific literature needs to speak very clearly to what the factors are and how to quantify them so that policymakers can understand them and react to that. In terms of specific policymaking or other things, I'll hesitate there because that's not my area of expertise. Where I'm coming from is that I consider it our job to get these investigations out into the literature and to promote the findings to policymakers who have the expertise to respond to and deal with these issues.

Dr. Zhang: I totally agree with Dr. Carton. I'll discuss a little bit more from a clinical care delivery perspective and the policy perspective. There were lots of efforts and policy efforts during the pandemic, either from the federal government, state government, or local government to address disparity directly and indirectly. There were some policies, not health policy, but other policies to help patients from racial and ethnic minority groups or from low socioeconomic groups to improve their health or improve their living status during the pandemic.

As we're entering a post-pandemic year, I think 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 disparities in the post-acute phase of COVID-19. And as we discussed, PASC in fact impacted the health and different organ systems of patients.

From the health system perspective, we should think about a patient who had COVID before they presented to the health system and the kind of resources providers would need to treat patients with PASC. PASC is not a single-symptom condition and it may impact multiple organ systems. So how does that inform planning and the allocation of health resources, especially to the regions that have a much higher prevalence or incidence of COVID-19? These regions may have a higher proportion of patients with PASC that will present to health systems. Are we ready to treat patients with PASC given the resources they have? As we know, COVID-19 impacted the employment of health care professionals. So, do we have a shortage of professionals to treat patients with PASC? I think these are very important PASC policy questions that need to be answered to help patients recover from COVID-19 infection, especially those from racial and ethnic minority groups.

Q. Have you found that Long COVID patients who were born or reside in large cities are worse off than patients from small cities?

Responses:

Dr. Zhang: For the first study we conducted about racial and ethnic disparities, most patients are in the New York City area, which is a very urban area. However, Dr. Carton's study really highlights some rural and urban differences, although we did not compare the PASC differences between rural and urban. But as Dr. Carton mentioned, Florida and New York City are two very different geographical regions. In New York City, we have very good public transportation and pretty good food access. But in Florida, the natural and social environment will be totally different. Many people have found that social characteristics can be risk factors for PASC. Considering these differences in environmental and social conditions, we would imagine there could be a difference in PASC symptoms between people from urban areas and people from rural areas. I think we'll be able to answer this question given the data that we have, which covers patients from different geographic locations in the US.

Dr. Carton: I agree. It's not a question that we directly posed of the data. As I mentioned, we grouped the environmental risk factors around an area defined by 5-digit ZIP Code and then identified the exposures to the patients within those regions. We didn't explicitly identify urban, peri-urban, and rural as variables of interest for these studies. Although, as Dr. Zhang mentioned, we can do so. I think to the degree to which these environmental risk factors vary from urban to peri-urban, suburban, and rural environments, we would be able to ask these questions. But in terms of the way we handled it for the study that Dr. Zhang and I described on exposome, we nested these variables within the ZIP Code of residence and didn't specifically interrogate by size of city.

Q. How do you check for inaccuracies in the data?

Responses:

Dr. Carton: Data quality is incredibly important. The fact that we have EHR data for a large number of patients and sites allows us to do a good amount of data quality and interrogation. I mentioned some of these checks that we have in place. We do quality checks as the data are moved into the repository, as I discussed in the earlier slides. Even within the confines of specific studies, we're able to interrogate issues of plausibility and completeness. For plausibility, are we seeing that breast cancer diagnoses among men would be a good example of implausible EHR data? For completeness, do we see extreme changes in a person's record over shorter periods of time?

There also are various other data quality checks that we're able to perform. For example, we can go back to the sites, present any data quality challenges to the contributing organizations, ask them to investigate, and to let us know if it's a problem with their source data or if it's something that was, for example, a mistake in the transmission.

The question was direct about how do we understand mistakes or inaccuracies in the coding? For very specific individual-level patient mistakes in diagnoses or coding, it's very difficult for us to catch, and it's a limitation of the EHR data. We're able to interrogate general patterns or quality issues by health system; some cases by site within system and some cases by specialty within system. But the data science that we have access to is not good enough to identify coding errors at the physician's office. This is case where we can identify general trends, but individual "mistakes" or sending back the data, we just don't have the ability to do that. It's more of a population science trend-based analysis.

Dr. Zhang: As Dr. Carton mentioned, there are some intrinsic limitations to the EHR data and we have done our best to address these limitations. The first I'll discuss is the accuracy of COVID-19 status. As you know, patients can get tested in a variety of places, such as in the health system or at the pharmacy, or elsewhere. This means that some patients appear COVID negative in EHR data, but they might have a positive test somewhere else that we don't see in the EHR data. This limitation can be mitigated somewhat by using data from multiple health systems.

which will allow us to do a much better job of capturing the results of COVID-19 tests compared with a study using EHR from a single health system. There are some things we can mitigate by using a more comprehensive dataset. But again, if a patient had a positive test outside of any EHR system, we just don't know; however, we can do something to mitigate this bias or inaccuracies.

Another topic more relevant to the first study we conducted is accuracy of race and ethnicity information in the EHR data. For people who use EHR data to understand racial and ethnic disparity, you may know that a high proportion of patients actually don't have race/ethnicity captured in the EHR data. We have a lot of patients where their race is listed as "unknown" or "other", or there is no information provided. For these patients, we really don't have accurate race/ethnicity information. It really poses a challenge for our study. However, there are some methods we can use to address this limitation. For example, natural language processing strategies can impute the missing or unknown race/ethnicity status based on clinical notes, which will require that we use unstructured data to capture patient race/ethnicity.

The point is that there are a lot of limitations in the EHR data. There are some things we can do to minimize the bias, but some limitations may still exist, even with the strategies and methods we've used to minimize these biases in the EHR data.

Q. What can patients do to help ensure accuracy in their own medical records?

Response:

Ms. Thomas: One thing to do is definitely use the 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.

Q. Are there similar exposome data found in other diseases, either viral or nonviral in origin, such as chronic cardiovascular disease, pulmonary disease, other diseases?

Response:

Dr. Zhang: Part of the motivation of this study is that many prior studies have reported that environmental factors, or what are called exposome factors, are risk factors for respiratory disease, such as chronic obstructive pulmonary disease (COPD) and some other conditions. Also, the built environment is important for patients with diabetes, heart failure, and some other conditions. There's very robust evidence on the association between exposome or environmental factors and a lot of chronic conditions. There's also very robust evidence on exposome or environmental risk factors and acuities of COVID-19, such as a rate of positive infection, hospitalization rate, and mortality. For example, a couple of studies have established an association between PM2.5 (fine particulate matter defined as particles 2.5 microns or less in diameter, which can be inhaled) and the mortality of COVID-19. All this gives us motivation to think there could be an association between exposome factors and Long COVID or PASC. I think that's part of the motivation as to why we pursued this study in the first place.

Q. 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?

Response:

Dr. Zhang: There are multiple pathways between race/ethnicity and PASC. Some of the pathways are direct and some are indirect associations, for example, mediated by some other factors. So, the question is, when we find a significant disparity by race/ethnicity, can it be purely because of race/ethnicity or can it be explained by other factors?

As I presented in the study, we adjusted for a set of covariates that could potentially be the confounders. Beyond that, what other factors could also explain racial/ethnic disparities? Based on some comments we received from the journal reviewers, we conducted additional analysis by adding some neighborhood-level socioeconomic factors, such as median income in ZIP Code area to see if socioeconomic status could further explain the race/ethnicity disparity. We found that after adding these economic factors, we still observed very significant disparities by race/ethnicity. That's why we call it an independent risk factor for PASC.

Q. By protein intake, are you referring to spike protein introduction via vaccination status or some other source?

Response:

Dr. Zhang: Dr. Carton and I are not experts on this question, though we do have some environmental health experts on our team. I think that protein part primarily means there are some biological disorders induced by the toxic air environment. I think we have at least one paragraph in our paper to discuss the potential mechanism behind the environmental spectrum and PASC. Feel free to check the article for more information.

Q. 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?

Response:

Dr. Carton: No, exposure to mold was not included. And we received some other questions in advance of the presentation related to indoor air pollution, which was also not measured. These were more of the external environmental factors that occur outside of the residence that we were able to gather from independent secondary data sources. That's just a limitation of the data that we have. However, there are other clinical cohort studies enrolling participants with consent and asking them direct questions about their behavior, behavioral history, environment, and other factors. However, I don't know if those studies are asking questions specific to mold. But there's an opportunity to study those specific questions related to individual exposures that's outside of both the environmental factors and the clinical factors that we can get through the EHR data and that are being investigated through other parts of RECOVER research.

Q. What is the rationale for adjusting for baseline comorbidities? Comorbidities cannot be confounders since they are not causes of race or racism. These are likely mediators, in which case statistical adjustment increases bias and masks disparities.

Response:

Dr. Zhang: There are certainly multiple pathways between race/ethnicity and PASC. By adjusting for comorbidities, we were able to test the direct associations between race/ethnicity and PASC. However, as you suggested, this may block the indirect associations, the associations mediated by comorbidities.

Webinar Slides

To request a copy of the R3 Seminar slides, please email RECOVER_ACC@rti.org.

To Learn More

- Information about RECOVER research and to volunteer for studies: https://recovercovid.org/research
- Frequently Asked Questions about RECOVER and PASC: https://recovercovid.org/faqs
- CDC information: Information for the general public and for healthcare providers about Post-COVID Conditions: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/
- For medical/scientific terminology: https://medlineplus.gov/healthtopics.html