Scientists Identify Characteristics to Better Define Long COVID

Using machine learning, researchers find patterns in electronic health record data to better identify those likely to have the condition.

May 17, 2022 By National Institutes of Health

A research team supported by the National Institutes of Health has identified characteristics of people with long COVID and those likely to have it. Scientists, using machine learning techniques, analyzed an unprecedented collection of electronic health records (EHRs) available for COVID-19 research to better identify who has long COVID.

Exploring de-identified EHR data in the National COVID Cohort Collaborative (N3C), a national, centralized public database led by NIH's National Center for Advancing Translational Sciences (NCATS), the team used the data to find more than 100,000 likely long COVID cases as of October 2021 (as of May 2022, the count is more than 200,000). The findings appear in The Lancet Digital Health.

Long COVID is marked by wide-ranging symptoms, including shortness of breath, fatigue, fever, headaches, "brain fog" and other neurological problems. Such symptoms can last for many months or longer after an initial COVID-19 diagnosis. One reason long COVID is difficult to identify is that many of its symptoms are similar to those of other diseases and conditions. A better characterization of long COVID could lead to improved diagnoses and new therapeutic approaches.

"It made sense to take advantage of modern data analysis tools and a unique big data resource like N3C, where many features of long COVID can be represented," said co-author Emily Pfaff, PhD, a clinical informaticist at the University of North Carolina at Chapel Hill.

The N3C data enclave currently includes information representing more than 13 million people nationwide, including nearly 5 million COVID-19-positive cases. The resource enables rapid research on emerging questions about COVID-19 vaccines, therapies, risk factors and health outcomes.

The new research is part of a related, larger trans-NIH initiative, Researching COVID to Enhance Recovery (RECOVER), which aims to improve the understanding of the long-term effects of COVID-19, called post-acute sequelae of SARS-CoV-2 infection (PASC). RECOVER will accurately identify people with PASC and develop approaches for its prevention and treatment. The program

also will answer critical research questions about the long-term effects of COVID through clinical trials, longitudinal observational studies, and more.

In the Lancet study, Pfaff, Melissa Haendel, PhD, at the University of Colorado Anschutz Medical Campus, and their colleagues examined patient demographics, health care use, diagnoses and medications in the health records of 97,995 adult COVID-19 patients in the N3C. They used this information, along with data on nearly 600 long COVID patients from three long COVID clinics, to create three machine learning models to identify long COVID patients.

In machine learning, scientists "train" computational methods to rapidly sift through large amounts of data to reveal new insights — in this case, about long COVID. The models looked for patterns in the data that could help researchers both understand patient characteristics and better identify individuals with the condition.

The models focused on identifying potential long COVID patients among three groups in the N3C database: All COVID-19 patients, patients hospitalized with COVID-19, and patients who had COVID-19 but were not hospitalized. The models proved to be accurate, as people identified as at risk for long COVID were similar to patients seen at long COVID clinics. The machine learning systems classified approximately 100,000 patients in the N3C database whose profiles were close matches to those with long COVID.

"Once you're able to determine who has long COVID in a large database of people, you can begin to ask questions about those people," said Josh Fessel, MD, PhD, senior clinical advisor at NCATS and a scientific program lead in RECOVER. "Was there something different about those people before they developed long COVID? Did they have certain risk factors? Was there something about how they were treated during acute COVID that might have increased or decreased their risk for long COVID?"

The models searched for common features, including new medications, doctor visits and new symptoms, in patients with a positive COVID diagnosis who were at least 90 days out from their acute infection. The models identified patients as having long COVID if they went to a long COVID clinic or demonstrated long COVID symptoms and likely had the condition but hadn't been diagnosed.

"We want to incorporate the new patterns we're seeing with the diagnosis code for COVID and include it in our models to try to improve their performance," said the University of Colorado's Haendel. "The models can learn from a greater variety of patients and become more accurate. We hope we can use our long COVID patient classifier for clinical trial recruitment."

This study was funded by NCATS, which contributed to the design, maintenance and security of the N3C Enclave, and the NIH RECOVER Initiative, supported by NIH OT2HL161847. RECOVER is coordinating, among others, the participant recruitment protocol to which this work contributes. The analyses were conducted with data and tools accessed through the NCATS N3C Data Enclave and supported by NCATS U24TR002306.

This news release was originally published by the National Institutes of Health on May 16, 2022.
