Machine learning identifies long COVID patterns from electronic health records

A machine learning algorithm identifies four reproducible clinical subphenotypes of long COVID from the electronic health records of patients with post-acute sequelae of SARS-CoV-2 infection within 30–180 days of infection; these patterns have implications for the treatment and management of long COVID.

This is a summary of:
Zhang, H. et al. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nat. Med.* https://doi.org/10.1038/s41591-022-02116-3 (2022).

Published online:
Published online: 13 January 2023

The question

Research has shown that patients with COVID-19 can suffer from a broad range of conditions and signs involving multiple organ systems after the acute phase of SARS-CoV-2 infection[1], including cardiovascular problems[2], metabolic problems[3] and neurological problems[4]. These studies support the existence of these sequelae as aspects of long COVID, but whether they are more likely to appear individually or co-appear in certain combinations is unclear. Our study aimed to fill in this knowledge gap by identifying patterns of coincident sequelae that occur within the post-acute infection period (30–180 days after a confirmed SARS-CoV-2 infection). For this analysis we used the electronic health records (EHRs) of patients from two large-scale clinical research networks in the US National Patient-Centered Clinical Research Network, INSIGHT and OneFlorida+, which include patients mainly from New York City and Florida, respectively.

The discovery

We created two cohorts of 20,881 patients from INSIGHT and 13,724 patients from OneFlorida+ who had a SARS-CoV-2 infection (confirmed by PCR or an antigen test) from March 2020 to November 2021 and also had at least one new post-acute sequela of SARS-CoV-2 infection from a list of 137 diagnostic categories curated from prior studies and clinical knowledge. We first tried to identify groups of post-acute sequelae of SARS-CoV-2 infection (PASC) using straightforward clustering methods, but these failed owing to the discrete nature and rarity or prevalence of the sequelae. We then drew an analogy to text analysis and developed an analytical methodology based on topic modeling.

Using this approach on the INSIGHT cohort, we identified four long COVID subphenotypes that corresponded to groups of patients with similar condition incidence patterns (Fig. 1). These subphenotypes were characterized by cardiac and renal conditions (subphenotype 1), respiratory, sleep and mood problems (subphenotype 2), musculoskeletal and neurological conditions (subphenotype 3), and digestive and respiratory system problems (subphenotype 4). Patients with subphenotype 1 were older (median age of 65 years, compared with 58 years overall), were more likely to be male (48.53%, compared with 41.63% overall), and were more likely to have been hospitalized or admitted to an intensive care unit (61.15% or 9.95%, compared with 43.37% or 5.48% overall, respectively) in the acute infection phase. Furthermore, the conditions in subphenotype 1 (such as heart and renal failure) were more severe, mostly with clear diagnostic criteria according to disease etiologies, and were more likely to be acute disease complications than the conditions in subphenotypes 2, 3 or 4 (such as dyspnea, musculoskeletal pain and abdominal pain), which were milder, mostly subjective to diagnose, and similar to patient-reported symptoms and signs. We validated these results by identifying the same four subphenotypes in the OneFlorida+ cohort, despite its different patient population and geographic region.

The implications

We identified four reproducible clinical subphenotypes of long COVID in SARS-CoV-2-infected patients with incident conditions within 30–180 days of infection, on the basis of their EHRs. We thus effectively 'teased out' patterns in the clinical heterogeneity of long COVID, which could potentially inform future research on the disease mechanisms of long COVID, treatment development and public health policymaking.

A limitation of our study is the use of EHRs, which record patients' information from their clinical visits. Consequently, the data we analyzed do not contain information from SARS-CoV-2-infected patients in these two clinical research networks who did not visit a clinic during the post-acute infection period, possibly owing to mild symptoms and signs, so we might have missed these patients. In addition, we analyzed the EHRs of patients who were infected with early SARS-CoV-2 variants, which meant that we did not assess the pattern of sequelae associated with the Omicron variant. Finally, our analysis is based on clinical observational data and thus cannot explain the disease pathophysiology of long COVID.

Our study provides evidence that the long COVID conditions do not appear independently but instead occur in patterns that correspond to reproducible subphenotypes. Follow-up studies are needed to further validate these subphenotypes in patient populations from other geographical areas in the United States and other countries. Data captured in more recent SARS-CoV-2 variant waves should also be analyzed to determine whether the subphenotype patterns have shifted. Basic science and translational researchers, clinical care professionals and public health policymakers should be aware of these long COVID clinical subphenotypes when making their action plans for managing the ongoing pandemic.

Fei Wang
Weill Cornell Medicine, New York, NY, USA.

EXPERT OPINION

"This study is not the first to cluster PASC features, but it is the largest and most robust I've seen (COVID negative comparison and validation in an independent cohort). Furthermore, the phenotypes identified are clinically interpretable and have corresponding medication prescription data as additional validation. The statistical approach is mostly well described and the data are comprehensive." **An anonymous reviewer.**

FIGURE

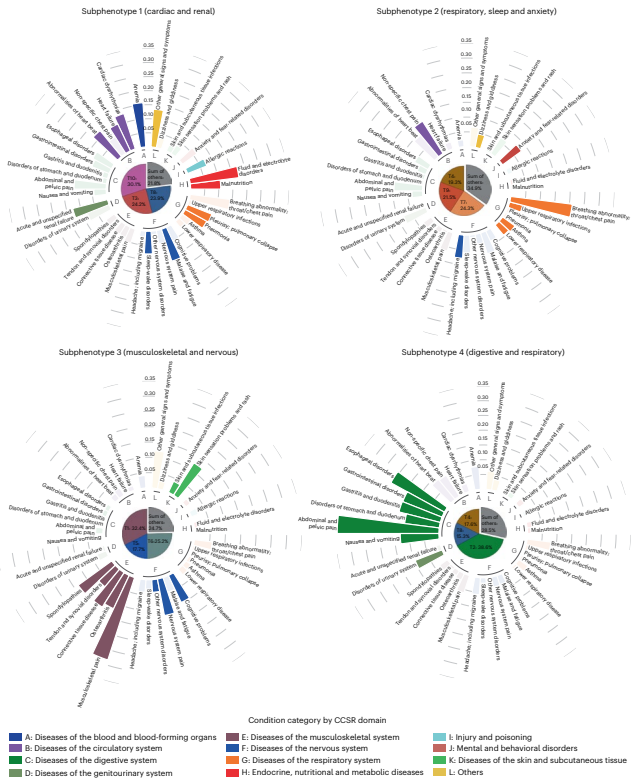

Fig. 1 | Long COVID subphenotypes identified in the INSIGHT cohort. Incident conditions within 30–180 days of SARS-CoV-2 infection were grouped into categories (A–K). Clustering analysis identified four subphenotypes involving coincident PASC. Colored bars indicate the prevalence of each condition (on a scale of 0–0.35 in terms of percentage). Center pie charts indicate the three most prevalent groups (topics T1–T10, which contain sets of conditions that tend to coincide), along with a fourth segment (gray) that is an aggregate of the remaining conditions.

BEHIND THE PAPER

Our study is part of the US National Institutes of Health RECOVER (Researching COVID to Enhance Recovery) initiative[5], which seeks to understand, treat and prevent PASC. The original idea stemmed from our clinicians' observations that patients with certain characteristics who were recovering from COVID-19 showed grouping of symptoms and signs in the post-acute infection period; for example, patients with autoimmune conditions tended to have neurological problems after acute SARS-CoV-2 infection. We therefore wondered whether we could detect such patterns in patient data. The insight to transition from conventional clustering methods to a statistical machine learning approach was crucial to identifying patterns in the EHR data. Our study is really a collaborative effort among clinicians, clinical-informaticians and data scientists. **F.W.**

REFERENCES

1. Al-Aly, Z., Xie, Y. & Bowe, B. High-dimensional characterization of post-acute sequelae of COVID-19. *Nature* **594**, 259–264 (2021).
   This article provides a systematic characterization of PASC.
2. Yan, X., Xu, E., Bowe, B. & Al-Aly, Z. Long-term cardiovascular outcomes of COVID-19. *Nat. Med.* **28**, 583–590 (2022).
   This article describes the range of cardiovascular disorders in the post-acute phase of SARS-CoV-2 infection.
3. Yan, X. & Al-Aly, Z. Risks and burdens of incident diabetes in long COVID: a cohort study. *Lancet Diabetes Endocrinol.* **10**, 311–321 (2022).
   This article describes the increased risk of diabetes in the post-acute phase of SARS-CoV-2 infection.
4. Evan, X., Xie, Y. & Al-Aly, Z. Long-term neurologic outcomes of COVID-19. *Nat. Med.* **28**, 2406–2415 (2022).
   This article reports the range of neurological disorders in the post-acute phase of SARS-CoV-2 infection.
5. RECOVER: Researching COVID to Enhance Recovery. https://recovercovid.org (2022).
   This website introduces and provides information about the RECOVER initiative.

FROM THE EDITOR

"Long COVID is a complex, heterogeneous condition that remains poorly understood, and few tools exist for the development of appropriate therapeutic strategies. Using EHRs and a machine learning approach, Wang et al. have identified four subphenotypes of long COVID that suggest stratified patient care pathways could be leveraged for the clinical management of long COVID." **Editorial Team, *Nature Medicine*.**