Viewpoint

Opportunities and Challenges in Using Electronic Health Record Systems to Study Postacute Sequelae of SARS-CoV-2 Infection: Insights From the NIH RECOVER Initiative

Hannah L Mandel[1], MS; Shruti N Shah[1], BS; L Charles Bailey[2], MD, PhD; Thomas Carton[3], PhD; Yu Chen[1], PhD; Shari Esquenazi-Karonika[1], PhD; Melissa Haendel[4], PhD; Mady Hornig[5], MA, MD; Rainu Kaushal[6], MD; Carlos R Oliveira[7,8], MD; Alice A Perlowski[9], MD; Emily Pfaff[10], PhD; Suchitra Rao[11], MD; Hanieh Razzaghi[2], MPH, PhD; Elle Seibert[12], BA; Gelise L Thomas[13], JD; Mark G Weiner[6], MD; Lorna E Thorpe[1], PhD; Jasmin Divers[14], PhD; RECOVER EHR Cohort[15]

[1]Department of Population Health, New York University Grossman School of Medicine, New York, NY, United States
[2]Applied Clinical Research Center, The Children's Hospital of Philadelphia, Philadelphia, PA, United States
[3]Louisiana Public Health Institute, New Orleans, LA, United States
[4]Department of Genetics, The University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, NC, United States
[5]Department of Epidemiology, Columbia University Mailman School of Public Health, New York, NY, United States
[6]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, United States
[7]Division of Infectious Diseases, Department of Pediatrics, Yale University School of Medicine, New Haven, CT, United States
[8]Division of Health Informatics, Department of Biostatistics, Yale University School of Public Health, New Haven, CT, United States
[9]Blooming Magnolia, Los Angeles, CA, United States
[10]Department of Medicine, The University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, NC, United States
[11]Department of Pediatrics, University of Colorado School of Medicine and Children's Hospital Colorado, Aurora, CO, United States
[12]Department of Neuroscience, USC Dornsife College of Letters, Arts and Sciences, Los Angeles, CA, United States
[13]Clinical and Translational Science Collaborative of Northern Ohio, Case Western Reserve University, Cleveland, OH, United States
[14]Department of Foundations of Medicine, New York University Long Island School of Medicine, Mineola, NY, United States
[15]See Acknowledgments,

Corresponding Author:
Hannah L Mandel, MS
Department of Population Health
New York University Grossman School of Medicine
180 Madison Avenue
New York, NY, 10016
United States
Phone: 1 732 314 1595
Email: Hannah.Mandel@nyulangone.org

Abstract

The benefits and challenges of electronic health records (EHRs) as data sources for clinical and epidemiologic research have been well described. However, several factors are important to consider when using EHR data to study novel, emerging, and multifaceted conditions such as postacute sequelae of SARS-CoV-2 infection or long COVID. In this article, we present opportunities and challenges of using EHR data to improve our understanding of long COVID, based on lessons learned from the National Institutes of Health (NIH)–funded RECOVER (REsearching COVID to Enhance Recovery) Initiative, and suggest steps to maximize the usefulness of EHR data when performing long COVID research.

(J Med Internet Res 2025;27:e59217) doi: 10.2196/59217

KEYWORDS

COVID-19; SARS-CoV-2; Long COVID, post-acute COVID-19 syndrome; electronic health records; machine learning; public health surveillance; post-infection syndrome; medical informatics; electronic medical record; electronic health record network;

electronic health record data; clinical research network; clinical data research network; common data model; digital health; infection; respiratory; infectious; epidemiological; pandemic

Introduction

Postacute sequelae of SARS-CoV-2 infection, colloquially known as long COVID, refers to ongoing, relapsing, new symptoms, or other health effects after the acute phase of SARS-CoV-2 infection (ie, present 4 or more weeks after the acute infection). Because long COVID is heterogeneous in presentation and may occur after a mild or even asymptomatic infection, it has posed unique and significant challenges for clinical research, including an overarching lack of data that are accessible and analyzable in a rigorous and reproducible manner to inform diagnostic criteria and care guidelines.

SARS-CoV-2 continues to impact populations globally with new and recurrent infections. While studies have provided a wide range of estimates regarding the proportion of patients with COVID-19 who develop long COVID [1], even conservative estimates point toward an enormous burden. Without established treatments targeting underlying pathophysiologic mechanisms, patients and clinicians have largely focused on treating symptoms [2] and managing organ damage, highlighting an urgent need to expand our knowledge of COVID-19 infection's long-term effects.

Electronic health records (EHRs) contain large quantities of clinical data that can be used for research without significant delay, making them an important data source for accelerated insight into emerging health conditions. However, challenges associated with leveraging EHR data for research can become amplified when mobilizing rapidly to study these conditions, particularly when they are multifaceted and lack clear hallmark traits or biomarkers.

The National Institutes of Health (NIH)–funded REsearching COVID to Enhance Recovery (RECOVER) Initiative was launched in 2021 [3] to advance research into long COVID. Broadly, RECOVER supports prospective clinical studies of adult and pediatric cohorts and real-world data studies that leverage EHR networks to study long COVID. These studies have been developing and validating algorithms to identify patients with long COVID within EHR networks for clinical and epidemiological characterization, risk factor prediction, and identification of prevention and treatment opportunities. RECOVER's real-world data efforts are led by 3 participating EHR research networks: the National COVID Cohort Collaborative (N3C); the National Patient-Centered Clinical Research Network (PCORnet); and PEDSnet, a pediatric learning health system within PCORnet. These networks are coordinated by a Clinical Science Core (CSC) at NYU Langone Health.

In this paper, we highlight opportunities and challenges of using EHR data to improve our understanding of long COVID, with a focus on analytical resources and scenarios we've encountered as US-based researchers, and conclude with some next steps for maximizing the usefulness of EHR data for long COVID research. Similar considerations may pertain to other chronic infection–associated diseases, that is, complex multisystemic conditions often occurring in the context of infection with other pathogens, including myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS).

EHR-Based Research Networks

Overview

The widespread adoption of heterogeneous health information systems across the United States hinders large-scale EHR-based research efforts. To overcome this, federal funding from the Patient-Centered Outcomes Research Institute (PCORI), Centers for Disease Control and Prevention (CDC), and NIH has facilitated the growth of networks that transform data into a common data model to facilitate aggregation across participating institutions. These networks are often referred to as clinical research networks (CRNs).

While such networks are a focus of this paper, we recognize that other successful international [4-7] and national models have been leveraged for large-scale long COVID studies. Domestically, the Department of Veterans Affairs (VA), which operates the largest integrated health care system in the United States, has made important contributions characterizing long COVID, its burden, and risk after reinfection, immunization or nirmatrelvir (Paxlovid) use [8-11]. The interoperability across many VA hospitals and clinics allows for data aggregation without the extensive harmonization or loss of data depth that usually accompanies a research network, though VA patients (military experience, mostly men) are not representative of the general population. Similarly, researchers have used claims data and outpatient laboratory results from UnitedHealth, a large health insurance provider, to characterize risk of long COVID in commercially insured patients [12]. The addition of claims data enables better understanding of health care use, inequities in access to care and costs associated with long COVID, though quantifying the larger economic burdens and disparities requires incorporation of data sources outside the health care system.

While researchers with access to EHR data from large clinical networks or health systems may bypass some challenges below, many national and international EHR-based collaborations wrestle with similar obstacles.

Opportunities

CRNs provide several benefits for conducting clinical research. Most significantly, they unify patient data that may otherwise be distributed across multiple health care facilities and facilitate access to large sets of timely, longitudinal patient information. This enables broad hypothesis generation, signal identification, and feasibility analysis, which is highly advantageous for emerging conditions like long COVID as such analyses can be conducted at a scale, pace, and cost that surpass traditional research methods [13]. Because CRNs use data generated during health care delivery, they are also well-positioned to study how practices evolve in ways that directly impact patients. For example, EHR data can be used to evaluate uptake and patterns of use for new diagnostic codes introduced to capture long

COVID [14], ME/CFS, and postural orthostatic tachycardia syndrome (POTS).

To overcome the lack of semantic interoperability across EHR systems and facilitate analysis, most CRNs map source data to common data schemas, such as the Observational Medical Outcomes Partnership (OMOP) [15] or PCORnet common data models [16], which provide frameworks for data standardization and are optimized for large-scale longitudinal data analysis. The popularity and extensibility of common data models helped CRNs to quickly adapt them for novel disease research at the beginning of the pandemic [17]. Use of a common data model also enables distributed network queries and analyses behind institutional firewalls, as well as centralization of the data for advanced artificial intelligence and machine learning [18].

In a relatively short time, a large amount of scientific literature has been generated regarding the clinical characterization and epidemiology of long COVID, yet efforts to synthesize this body of research have been hampered by the inconsistency of methodologic approaches, study definitions, and findings. Research conducted using a shared infrastructure or data model promotes reproducible research and exchange of clinical and variable definitions that can be harmonized and tested across multiple health systems. For example, a recent collaboration between RECOVER and NIH's All of Us study demonstrated success in reusing long COVID definitions across OMOP data environments, enabling reproduction of N3C's definition within the All of Us database [19].

Along the same lines, CRNs are often well-positioned to foster multidisciplinary collaboration [20]. They have streamlined contracting and collaboration processes that expedite research while reducing regulatory burden, and can also provide an efficient vehicle for patient recruitment into prospective cohort studies and clinical trials.

Challenges

Some broad challenges come with using CRNs for population health research and may directly impact research into long COVID, where data quality, depth, and timeliness are crucial. These largely arise from the need to standardize data across multiple institutions, which may result in a loss of detail and context from the originating EHR [21]. For example, while the OMOP and PCORnet data models support documentation of patient cause of death and discharge disposition, not all do; furthermore, not all sites collect and report these data. Analyses can also be limited by not having information such as provider specialty, medication indication, or drug dosing consistently abstracted from the EHR. However, as described earlier, the extensibility of many common data models means that new concepts can be incorporated to close priority gaps.

Aggregating data within a CRN can mask important heterogeneity due to differences in how health systems and individual practitioners provide and document care. One example is the variation in how visit data are represented across individual health care facilities. The definition of an "encounter" in a given health system's data has more to do with business rules than clinical care; for example, some systems create a separate encounter record for each day of an inpatient stay, while others use a single record to capture the entire hospitalization. As most variables in EHR data are captured at the visit level, inconsistency in the definition of encounters can lead to errors when analyzing clinical observations. Thus, to accurately classify types of visits and to quantify visits across sites, careful standardization is required [22]. Ongoing governance, documentation, and input from data experts at contributing institutions are essential to bridge syntactic and semantic differences [21].

Timeliness of data delivery represents an additional challenge. While data from CRNs are generally rapidly available, there is still a latency from participating institutions. Data must be mapped, transformed, transmitted, quality checked, and potentially resubmitted with corrections, which takes time and collaborative effort. Delays may also result from the need to incorporate data standards for new concepts (such as the long COVID *ICD-10* [*International Statistical Classification of Diseases, Tenth Revision*] code U09.9, which was not introduced until late 2021).

Finally, real-world data present specific limitations compared with traditional research data. These include data missingness, which has been described extensively [23] and can exist for various reasons (eg, patient discontinuation of care or data points not being systematically captured in EHRs). EHR data are also subject to biases like selection bias [24], where patients accessing care may be sicker, more likely to be insured, and able to take time off work and afford transportation [13,25]. Data from nonacademic institutions, and other settings that may lack robust data systems such as nursing homes, are often underrepresented in CRNs. Careful study designs, informatics-informed approaches to transforming raw clinical data into analytical datasets [26], techniques to supplement structured EHR data with items extracted from free text notes such as social determinants [27], and advanced statistical approaches are often required to minimize bias.

Identifying Patients With Long COVID Within EHRs

Overview

Robust case definitions are essential to disease surveillance, enabling consistent use of data from collection through analysis and across population subgroups, time, and geographic jurisdictions. However, Long COVID still lacks a uniform case definition, and multiple frameworks have emerged with varying timeframes and symptoms [25]. This is partly due to the heterogeneity of long COVID's clinical presentation, which includes substantial differences between adults and children [28] and hundreds of potential signs and symptoms. The National Academies of Sciences, Engineering, and Medicine's recent report A Long COVID Definition provides updated recommendations around diagnosis, noting that long COVID can impact any organ system; follow acute infection of any severity, including asymptomatic infection; be continuous from acute infection or delayed in onset; and present as incident or worsening, mild or severe, and persistent or intermittent [29]. These broad guidelines leave room for ambiguity, and as there

is no diagnostic test for long COVID, their application at the point of care involves clinical judgement, experience, and thorough differential diagnosis to establish probable causality.

Finally, we are still learning about long COVID's disease course and trajectory, how it relates to development of clinical conditions such as diabetes, kidney disease, and POTS, and how its characteristics may evolve over time. While these complexities underlie all long COVID research, EHR-based research is particularly well-positioned to inform and validate potential definitions.

Opportunities

Researchers working with EHR data operationalize case definitions by creating computable phenotypes, which algorithmically identify cohorts of interest within an EHR system [30]. Generating, adapting, and validating computable phenotypes can be complex, even for well-characterized diseases and conditions, so multiple approaches for identifying patients with long COVID have evolved both nationally and internationally. These include building computable phenotypes using "rules-based" approaches or prespecified logic-based queries to search for specific criteria such as diagnosis codes, medications, or laboratory results based on clinical literature and expertise [31-33], and data-driven or inductive approaches, including machine learning (ML) to automatically classify long COVID without detailed prespecification of the inclusion or exclusion criteria [34,35]. These approaches can also be combined, leveraging insights from ML to shape rules-based definitions.

The lack of a formal case definition, limited SARS-CoV-2 testing information, and large amounts of rich longitudinal data within EHRs make ML approaches particularly attractive. Supervised ML models can be trained on sets of patients that are certain or highly likely to have long COVID, such as patients assigned a long COVID diagnostic code, and then used to identify patients with probable long COVID who may not have a documented or readily recognized diagnosis [35]. ML and other statistical approaches have also been applied to recognize incident diagnoses that appear more frequently in cohorts with a recent history of COVID-19 infection than those without [36,37].

Although the EHR does not capture every case (or every recurrence) of acute SARS-CoV-2 among its patients, there is still significant opportunity to use EHRs to study the potential effect of SARS-CoV-2 reinfection on risk of long COVID among those patients with multiple recorded cases. Hadley et al [38] found that across SARS-CoV-2 variants, the incidence of new long COVID diagnoses after reinfection is lower than after the initial SARS-CoV-2 infection. Bowe et al [11] found that though risk is lower with more recent variants, it is cumulative across infections. As multiple reinfections become more common, there will be continual opportunities to reevaluate these findings in the context of new variants and subvariants.

Both rules-based and ML techniques are being used to identify potential clusters or subphenotypes of long COVID [39-41]. For example, an analysis by Zhang et al [41] identified four distinct subphenotypes (including cardiac and renal, musculoskeletal, and nervous system) and specific risk predictors for each. Clustering analyses may uncover diverse symptoms that have common biological causes, and in establishing distinct groupings around them, facilitate diagnosis and treatment targeted toward specific subtypes [42].

Challenges

Within the EHR, Long COVID may not be well-documented due to diagnostic complexities, as described earlier. The initial focus among clinicians on respiratory symptoms and widespread lack of knowledge around what were initially considered to be "atypical" presentations, such as postexertional malaise, ME/CFS, POTS, and dysautonomia, also contribute to missingness of relevant diagnoses within the EHR [43], which is compounded by the absence of or delay in introduction of *ICD-10* codes for many of these conditions, including the U09.9 code for long COVID itself**.** Such manifestations can significantly impact daily function but may not be considered attributable to previous COVID-19 infection (or misdiagnosed as mental health issues), leading to their underrepresentation in research [43].

Health care professionals may be subject to implicit biases that impact documentation of patient information (eg, social determinants and symptoms) within the EHR [24]. This may also lead to underestimation of long COVID for certain populations, such as racial or ethnic minorities. Indeed, U09.9 diagnostic codes have been found to be more common among patients that are non-Hispanic White and living in areas of low poverty compared with the overall COVID-19–positive population, potentially reflecting differences in access to care and other interactions with the health care system [14]. Although we found value in using visits to a long COVID clinic as an early proxy for a long COVID diagnosis code [35], both of these markers have biases and should not be considered a gold standard for long COVID identification [44]. There is also substantial need to complement patient-reported symptoms with clinically characterized ones [45] so that physicians are not "gatekeepers" of documented symptoms.

Given that long COVID is likely underdiagnosed and underdocumented, computable phenotype algorithms must supplement formal U09.9 diagnoses with other patient data following acute SARS-CoV-2 infection. However, identifying acute infections is also not straightforward. Diagnosis codes such as the ICD-10 code U07.1 have demonstrated variable sensitivity and positive predictive value [46,47], which may be exacerbated by "rule out" diagnoses or improper coding. Even when diagnoses are correct, the true timing of the initial infection with respect to the date of the diagnosis code may be uncertain [44], and we found this to complicate EHR-based calculations of long COVID incidence. Limited availability of SARS-CoV-2 testing during the early phase of the pandemic, widespread testing at nonclinical settings (eg, kiosks and pharmacies), and the later popularity of at-home tests have all contributed to low confidence in accurately capturing SARS-CoV-2 infection (and reinfection) history within the EHR, and amplified potential misclassification of long COVID patients without documented acute COVID-19 infection. We

note, however, that misclassification of patients as uninfected is an issue that extends beyond EHR-based studies [48], broadly impacting long COVID research.

ML methods are also subject to obstacles, including the lack of a validated test set for long COVID and dependency of results on the training set used, which may limit reproducibility (although, as demonstrated by the collaboration between RECOVER and All of Us, ML algorithms can be replicated in new environments) [19]. ML models must be interpretable, so predictions can be explained instead of only existing within a "black box."

Determining Risk Factors for Long COVID Using EHRs

Risk factors for long COVID include demographic characteristics, lifestyle behaviors, underlying comorbidities, and social and environmental factors. Identifying risk factors can guide screening and prevention efforts, particularly in highly susceptible groups, and potentially facilitate tailored long COVID treatments.

Opportunities

EHR-based studies are primed to quickly reveal possible associations and risk factors for conditions like long COVID due to large volumes of available data and ability to stratify by various subgroups. For example, Rao et al [28] leveraged PEDSnet's geographical and clinical breadth to identify risk factors such as age, severity of acute COVID-19 infection, and comorbid conditions. Potential biological variables that are well-defined in EHRs and therefore, more reliably available for application in adjusted analyses include age, sex, BMI, and blood pressure, and, indeed, these variables have been shown to confound the evaluation of causal pathways for long COVID [49-51].

Where data of interest are missing from the EHR, linkage to external sources can increase data quality and completeness. Statewide vaccination or vital record data can be acquired and linked at the patient or area level to enhance the accuracy of vaccine status and COVID-19 outcomes [52,53]. Linkage with geographic information systems (GIS) can address questions about complex networks of long COVID burden that pertain to certain geographic regions, including the so-called "exposome" [54] even down to ZIP codes and neighborhoods [13,14]. Patient-reported outcomes from patient portals and surveys can improve the completeness of information on demographic, social, and behavioral factors such as race, ethnicity, substance use, exercise frequency, and smoking status [13,55]. To enhance completeness of data such as outpatient medication use, linkage to claims data can be an effective strategy. Data of interest may also be captured within free-text clinical notes, images, and other scanned documents, requiring application of natural language processing tools, including cutting-edge language models [27], or manual chart review to incorporate them.

Within EHR-based studies, several analytic approaches can address confounding and effect modification and help measure mediating influences. Researchers can match comparison groups using various approaches, including propensity scores for known confounders. In addition, causal mediation analysis methods may more accurately explain observed associations between risk factors and long COVID development.

Designing studies with appropriate comparison groups for analyses of risk factors can be challenging. If comparison groups are matched too closely, relevant differences in concurrent risk factors and outcomes may be masked. If groups are not matched well enough, then the impact of confounders may overwhelm the ability to detect differences associated with the characteristic of interest. Researchers using real-world data can conduct sensitivity analyses under different assumptions of risks and outcomes. For example, Hill et al [56] leveraged 3 different control groups and 2 definitions of long COVID to enhance an investigation of risk factors within N3C. This agility is especially useful with a novel condition such as long COVID, where our understanding is rapidly evolving. Comparison groups can be iteratively modified, and analyses rerun as new information or criteria emerge.

EHR data, especially within the context of a CRN, also facilitates the study of risk factors within rare populations and subgroups (eg, peoples indigenous to the United States). Such groups often experience a disproportionate burden of COVID-19 complications due to socioeconomic and structural factors [51,57], but are generally not well-represented in prospective studies and randomized controlled trials [58]. EHR data can similarly help overcome the methodological constraints and recruitment challenges of rare disease research [13], providing an opportunity to study rare conditions and outcomes and facilitating clinical trial cohort selection.

Finally, EHRs provide rich datasets for pragmatic clinical trials and target trial emulation to identify potential therapies [59,60]. For example, there is inconclusive data on the efficacy of Paxlovid for preventing or treating long COVID [61]; novel approaches such as target trial emulation using EHR data can help evaluate its efficacy and determine which patients should be prioritized for treatment based on risk factors for severe long COVID subphenotypes. The ability to quickly evaluate and personalize potential treatments is highly advantageous given the large burden of long COVID and its heterogeneous presentation. Because target trial emulation better reflects real-world medication adherence (or application of other interventions), comparing the results of emulated and clinical trials, and examining how the methods and populations differ, may yield insights into generalizability of clinical trial results and sources of bias within observational analyses.

Challenges

Limitations outlined earlier, such as data missingness, selection bias, and implicit biases, can directly impact the ability to evaluate risk factors using EHR-based analyses. Potential risk factors can be rigorously evaluated if well-defined and well-captured, but many important risk factors may not be completely, meaningfully, or consistently recorded within the EHR.

For example, race and ethnicity are important demographic risk factors for long COVID [51], but these data can be missing in

over 25% of patient records [55,62]. Lack of robust, structured data on social and behavioral determinants of health, such as physical activity levels, smoking status, and alcohol use [13,51,62] may necessitate supplementation with survey results or data extracted from clinical notes as described above. EHR-based assessments of symptom severity and improvement may also be of limited use without these approaches.

Imperfect data capture also impacts our ability to establish proper comparison groups. This notably includes the shrinking population of patients without previous SARS-CoV-2 infection [63]. In addition to being challenging to identify, control patients identified by negative lab tests may actually be in worse health than those with mild COVID-19 infection, likely due to routine testing of patients within the emergency department or inpatient settings. Another example applies to our work examining obstructive sleep apnea (OSA) as a risk factor for long COVID [64]. A comparison group of patients ostensibly without OSA may include some with undiagnosed, subclinical, or undocumented OSA, and sizable numbers of misclassified patients would lead to underestimation of effect size on long COVID. Similarly, slight improvements in recognition of ME/CFS throughout the pandemic may have resulted in some individuals with prepandemic long COVID–like symptoms (eg, unrefreshing sleep and postexertional malaise) having their illness misclassified as SARS-CoV-2–related. Appropriate comparison groups must be carefully considered.

The pandemic disrupted typical health care use and documentation patterns. Clinic closures and reduced access to specialty care may have masked diagnoses of chronic conditions and comorbidities and led some patients with long COVID, particularly those with milder presentations, not to seek care. This biases EHR data, especially from the first year of the pandemic, toward underascertainment of health issues; prepandemic controls (who we can say with certainty have not had COVID-19 infection) may have higher rates of clinical events than pandemic-era patients. Similarly, research about the effectiveness of vaccinations in reducing the risk of long COVID is hampered by widespread distribution in nonclinical settings [65] and reinforces the importance of ongoing data exchange between EHRs and vaccination registries in addition to dedicated linkage efforts.

These challenges also apply to potential confounding, which is particularly important within observational research on risk factors and causal mechanistic pathways for long COVID. Certain potential confounders, such as age or underlying health conditions, are easier to adjust for than others. The severity of acute COVID-19 infection, which appears in some studies to be directly associated with the risk of developing long COVID [49,50,66], is a particularly illustrative example. More severe COVID-19 infection, whether it results in hospitalization or not, may result in increased frequency of diagnosis and monitoring of long COVID and higher data density for affected patients within the EHR. These biases may contribute to the underrepresentation of asymptomatic, mild, or moderate acute COVID-19 infections, as well as greater clinician awareness of long COVID in the context of severe illness (although most cases occur among nonhospitalized patients) [43]. Analyses stratified by acute illness severity are important to differentiate between nonhospitalized and hospitalized patients.

Discussion

As SARS-CoV-2 and our susceptibility to infection evolve, sustained research efforts are needed to build our knowledge of long COVID and evaluate its presentation over time. Within a short timeframe, the use of EHR networks to study long COVID has grown rapidly, nationally and internationally. The immediate need to advance our understanding of long COVID must be balanced with a thoughtful methodological approach that leverages the many strengths of EHR data while recognizing their limitations. EHRs have been used to characterize long COVID and its subphenotypes [14,32,40,41,45], estimate its burden [28,67], identify risk factors [49-51,66], and evaluate SARS-CoV-2 vaccine efficacy for long COVID prevention [8,53,68]. Although we anticipate that the breadth and rigor of this research will increase as computable phenotypes improve through iterative validation and data linkage becomes more extensive, several innovations to effectively wield EHR data, such as improved accuracy and consistency of recording of long COVID-specific diagnoses, application of ML to identify patients with long COVID and harmonized data analyses across CRNs, have already begun increasing the capacity of investigators to study long COVID and, importantly, augment research on long COVID subphenotypes.

EHRs can be instrumental in supplementing randomized controlled trials to evaluate treatments and preventative strategies such as vaccinations, but novel approaches such as target trial emulation have not yet widely extended to long COVID. Increased data linkage to vaccination registries, survey data, and other sources can further enrich EHR-based analyses, providing more complete patient outcomes and medical history data. As recruitment into prospective RECOVER studies increases, efforts are underway to link EHR data with participant study data, enhancing both datasets for validation and analysis.

We envision a near future where the mutual benefits of clinical and real-world data research collaborations have been optimized for synergistic learning. EHRs can identify study participants and generate hypotheses that can be prospectively tested. They can also inform clinical trial feasibility, generate preliminary data around specific populations, therapeutics, and outcomes of interest, and inform the interpretation of clinical trial results through trial emulation. However, strengthening this interchange will require new workflows, collaboration, and time.

We caution that despite the dedication of the research community and high expectations for EHRs to quickly advance our understanding of long COVID, sustained, collaborative efforts are needed for this work. New treatments and vaccines are still being introduced, requiring process modifications for extraction and standardization of data. Complex clinical traits require careful development of computable phenotypes and rigorous data quality evaluation to identify specific misclassification risks. Continued coordination is necessary to converge upon reliable, repeatable insights and yield effective strategies for long COVID.

Acknowledgments

This study is part of the NIH RECOVER (National Institutes of Health's REsearching COVID to Enhance Recovery) Initiative, which seeks to understand, treat, and prevent the postacute sequelae of SARS-CoV-2 infection. For more information please visit the website [69]. This research was funded by an NIH Agreement (OTA OT2HL161847) as part of the RECOVER research program. We would like to thank the National Community Engagement Group (NCEG), all patient, caregiver, and community representatives, and all the participants enrolled in the RECOVER Initiative.

We would also like to thank collaborators in the RECOVER Consortium[discontinued effort on RECOVER].

National COVID Cohort Collaborative (N3C):

Axle Informatics: Kristen Hansen. Berkeley Lab: Justin Reese. Columbia University: Karthik Natarajan. Datavant: Jasmin Phua. Duke University: Warren Kibbe, PI. Emory University: Richard Moffitt, PI, Margaret Hall, Rishi Kamaleswaran, Jason Yoo. Endeavor Health: Anthony Solomonides, PI. Illinois State University: Nariman Ammar. Jackson Labs: Hannah Blau, Peter Robinson. Johns Hopkins University: Christopher G. Chute, PI, Ali Afshar, G. Caleb Alexander, Lisa Eskenazi, Tricia Francis, Davera Gabriel, Kirby Gong, Stephanie Hong, Harold Lehmann, Hemal Mehta, Chirag Parikh, Ann Parker, Rayna Xiao, Tanner Zhang, Richard Zhu, Jared Zook. Minderoo Center for Federated Cancer Research: Robert Miller. National Institutes of Health: Hythem Sidky. Northeastern University: Christian Reich, PI, Kristin Kostka. Oregon Community Health Information Network (OCHIN): Brenda McGrath, PI, Treasure Allen, Rob Schuff. Oregon Health & Science University: Erik Benton, David A. Dorr, Justin Ramsdill. Palantir Technologies: Maya Choudhury, Andrew Girvin. Patient Lead Research Collaborative: Hannah Davis, PI, Gina Assaf, Lisa McCorkell, Yochai Re'em, Anisha Sekar, Hannah Wei. RTI International: Daniel Brannock, Rob Chew, Emily Hadley, Alexander Preiss. Scripps: Ginger Tsueng. State University of New York at Stony Brook: Janos Hajagos, PI, Rachel Wong, PI, Adit Anand, Raj Gupta, Mengyao Hu, Prahathish Kameswaran, Saarthak Kapse, Eileen Keck, Farrukh koraishy, Spencer Krichevsky, Saaya Patel, Joel Saltz, Mary Saltz, Hiteshwar Singh, Shreya Sinha, Sam Soff, Kimon Stathakos, Chelsea Twan, Rohith Vaddavalli, Sai Rachana Yerram. The National Institute of Diabetes and Digestive and Kidney Diseases: Kenneth J. Wilkins. TriNetX: Lora Lingrey, Matvey Palchuk. Trinity Health: Joe Flack. Tufts Medical Center: Andrew Williams, PI. University of California Davis: Konstantin Kunze, PI. University of Chicago: Tom Best, PI. University of Iowa: Dave Eichmann, PI, Charisse Madlock. University of Kansas Medical Center: Kelechi Anuforo, PI. University of Kentucky: Ramakanth Kavuluru. University of Maryland Baltimore: Stacy Dalton, PI. University of Massachusetts: Feifan Liu, PI. University of Michigan: J. Brian Byrd. University of Minnesota: Steve Johnson, PI, Ashley Benner, Carolyn Bramante, Scott Chapman, Duy Duong, Michael Evans, Jared Huling, Corey McGee, Zheng Wang, Talia Wiggen, Rui Zhang. University of Nebraska Medical Center: Jerrod Anzalone. University of North Carolina at Chapel Hill: Emily Pfaff, PI, Melissa Haendel, PI, Abhishek Bhatia, Sofia Dard, Liz Kelly, Peter Leese, Tomas McIntee, JP Powers, Bryan Laraway, Julie McMurry, Shawn T. O'Neil, Chris Roeder, Anita Walden. University of Pittsburgh: Michele Morris. University of Rochester: Elaine Hill, PI, Jack Chang, Adam Dziorny, Daniel Guth, Tanzy Love, Klint Mane, Sharad Kumar Singh, Richa Yadav, Ayushi Saxena. University of Texas Medical Branch: Heidi Spratt, PI. University of Utah: Jackson Barlocker. University of Virginia: Don Brown, PI, Sihang Jiang, Johanna Loomba, Saurav Sengupta, Suchetha Sharma, Andrea Zhou. University of Washington: Rena Patel. University of Texas Health Science Center at Houston: Hongfang Liu, PI. Virginia Commonwealth University: Brian Bush, PI.

PCORnet Core Contributors:

Ann & Robert H. Lurie Children's Hospital of Chicago: Ravi Jhaveri. Children's Hospital of Philadelphia: L. Charles Bailey, PI, Christopher B. Forrest, PI, Andrea Allen, Rodrigo Azuero-Dajud, Andrew Samuel Boss, Morgan Botdorf, Colleen Byrne, Peter Camacho, Abigail Case, Kimberley Dickinson, Susan Hague, Jonathan Harvell, Miranda Higginbotham, Kathryn Hirabayshi, Sandra Ilunga, Rochelle Jordan, Kyle Kays, Aqsa Khan, Vitaly Lorman, Nicole Marchesani, Sahal Master, Jill McDonald, Nhat Nguyen, Hanieh Razzaghi, Qiwei Shen, Alexander Shorrock, Levon H. Utidjian, Ryan Webb, Kaleigh Wieand. Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine: Nate M. Pajor. Harvard Pilgram: Jason Block. Louisiana Public Health Institute: Tom Carton, PI, Anna Legrand, Elizabeth Nauman. Nationwide Children's Hospital and The Ohio State University College of Medicine: Jennifer Muszynski. Nemours/Alfred I. duPont Hospital for Children: H. Timothy Bunnell, Christopher Pennington, Cara Reedy. The Perelman School of Medicine, University of Pennsylvania: Yong Chen. University of Colorado School of Medicine and Children's Hospital Colorado: Suchitra Rao, PI. Vanderbilt University Medical Center: Russell Rothman, PI. Weill Cornell Medicine: Rainu Kaushal, PI, Sajjad Abedian, Dominique Brown, Christopher Cameron, Thomas Campion, Andrea Cohen, Marietou Dione, Rosie Ferris, Wilson Jacobs, Michael Koropsak, Alex LaMar, Colby V. Lewis, Dmitry Morozyuk, Peter Morrisey, Duncan Orlander, Jyotishman Pathak, Mahfuza Sabiha, Edward J. Schenck, Stephenson Strobel, Zoe Verzani, Fei Wang, Mark Weiner, Zhenxing Xu, Chengxi Zang, Yongkang Zhang.

Observational Consortium Steering Committee:

Jeffrey Burns, Co-Chair; Serena Spudich, Co-Chair; Charles Bailey, Mine Cicek, Melissa M. Cortez, Felicia Davis Blakley, Andrea S. Foulkes, David Goff, Stuart D. Katz, Jessica Lasky-Su, Torri D. Metz, Lisa T. Newman, Igho Ofotokun, Sudha Seshadri, Melissa Stockwell, James Stone, Brittany D. Taylor, PJ Utz, Neely A. Williams.

Administrative Coordinating Center at Research Triangle Institute International:

Lisa T. Newman, PI, Julie Abella, Quinn Barnette, Christine Bevc, Jennifer Beverly, Patricia Ceger, Julie Croxford, Emily Cunningham, Mike Enger, Katie Fain, Tonya Farris, Sean Hanlon, David Hines, Vicki Johnson-Lawrence, Kevin Jordan, Craig Lefebvre, Beth Linas, Bryan Luukinen, Meisha Mandal, Nikki J. McKoy, Susan Nance, Ashleigh Oakland, Demian Pasquarelli, Claire Quiner, Rita Sembajwe, Gwendolyn Shaw, Vanessa Thornburg, Kendall Tosco, Hannah Wright.

Clinical Science Core at NYU Langone Health:

Rachel S. Gross, PI; Judith S. Hochman, PI; Leora I. Horwitz, PI; Stuart D. Katz, PI; Andrea B. Troxel, PI; Lenard Adler, Precious Akinbo, Ramona Almenana, Malate Aschalew, Lara Balick, Ola Bello, Sultana Bhuiyan, Nina Blachman, Ryan Branski, Jasmine Briscoe, Shari Brosnahan, Elliott Bueler, Yvette Burgos, Nina Caplin, Domonique Nicole Chaplin, Yu Chen, Shen Cheng, Peter Choe, Jess Choi, Alicia Chung, Richard Church, Stanley Cobos, Nakia Croft, Angelique Cruz Irving, Phoebe Del Boccio, Iván Díaz, Jasmin Divers, Vishal Doshi, Benard Dreyer, Samantha Ebel, Shari Esquenazi-Karonika, Arline Faustin, Elias Febres, Jeffrey Fine, Sandra Fink, Catherine Freeland, Jennifer Frontera, Richard Gallagher, Alejandra Gonzalez-Duarte, Denise Hasson, Sophia Hill, Jennifer Hossain, Shahidul Islam, Stephen Johnson, Neha Kansal, Rachel Kenney, Tammy Kershner, Deepshikha Kewlani, Judy Kwak, Michelle F. Lamendola-Essel, Sarah Laury, Gregory Laynor, Lei Lei, Terry Leon, Zoe A. Lewczak, Janelle Linton, Max Logan, Nadia Malik, Lia Mamistvalova, Hannah Mandel, Gabrielle Maranga, Patenne Dolores Mathews, Aprajita Mattoo, Tony Mei, Alan Mendelsohn, Emmanuelle Mercier, Patricio Millar Vernetti, Marc Miller, Maika Mitchell, Andre Moreira, Praveen C. Mudumbi, Erica Nahin, Nandini Nair, Joseph Nekulak, Kellie Owens, Brendan Parent, Nandan Patibandla, Peter Petrov, Radu Postelnicu, Francesca Pratt, Isabelle Randall, Priyatha Rao, Amy Rapkiewicz, JohnRoss Rizzo, Johana Rosas, Chelsea Rose, Christina Saint-Jean, Michelle Santacatterina, Binita Shah, Aasma Shaukat, Naomi Simon, Aylin Simsir, Miranda Stinson, Wenfei Tang, Vasishta Tatapudi, Sujata Thawani, Mary Thomas, Lorna Thorpe, MeeLee Tom, Ethan Treiha, Jennifer Truong, Mmekom Udosen, Carlos Valencia, Jessica Velazquez-Perez, Patricio Millar Vernetti, Crystal Vidal, Anand Viswanathan, Amy Willerford, Natasha Williams, Crystal Wong, Marion J. Wood, Shannon W. Wuller, Shonna H. Yin, Chloe Young, Jonah Zaretsky, Susanna Zavlunova

Authors' Contributions

HLM, SNS, LCB, JD, EP, HR, MGW, and LET contributed to conceptualization. HLM, SNS, LET wrote the original draft of this manuscript. LCB, TC, YC, JD, SE-K, MH, CRO, EP, SR, HR, MGW, MH, EH, GLT, AAP, and RK handled reviewing and editing of this manuscript. SE-K contributed to project administration.

Conflicts of Interest

LCB, TC, EP, HR, and MW report funding from the Patient-Centered Outcomes Research Institute. HR, LCB, TC, EP, and MW report research funding from the NIH. LCB reports research funding from the National Institute of Diabetes and Digestive and Kidney Diseases, and Centers for Disease Control and Prevention. CO has received grants from the National Institutes of Allergy and Infectious Diseases, has a council role with the Eastern Society of Pediatric Research, and received an honorarium from the NIH. EP reports consulting fees from the NIH and a speaker honorarium from Cincinnati Children’s Hospital. SR has received consulting fees from Sequiris. All other authors declare no competing interests.

References

1. Woodrow M, Carey C, Ziauddeen N, Thomas R, Akrami A, Lutje V, et al. Systematic review of the prevalence of long COVID. Open Forum Infect Dis. 2023;10(7):ofad233. [FREE Full text] [doi: 10.1093/ofid/ofad233] [Medline: 37404951]
2. Brown K, Yahyouche A, Haroon S, Camaradou J, Turner G. Long COVID and self-management. Lancet. 2022;399(10322):355. [FREE Full text] [doi: 10.1016/S0140-6736(21)02798-7] [Medline: 35065779]
3. RECOVER: researching COVID to enhance recovery. Recover. URL: https://recovercovid.org [accessed 2025-01-16]
4. Thygesen JH, Tomlinson C, Hollings S, Mizani MA, Handy A, Akbari A, et al. Longitudinal Health and Wellbeing COVID-19 National Core Study and the CVD-COVID-UK/COVID-IMPACT Consortium. COVID-19 trajectories among 57 million adults in England: a cohort study using electronic health records. Lancet Digit Health. 2022;4(7):e542-e557. [FREE Full text] [doi: 10.1016/S2589-7500(22)00091-7] [Medline: 35690576]
5. Lam ICH, Wong CKH, Zhang R, Chui CSL, Lai FTT, Li X, et al. Long-term post-acute sequelae of COVID-19 infection: a retrospective, multi-database cohort study in Hong Kong and the UK. EClinicalMedicine. 2023;60:102000. [FREE Full text] [doi: 10.1016/j.eclinm.2023.102000] [Medline: 37197226]
6. Magnusson K, Turkiewicz A, Flottorp SA, Englund M. Prevalence of long COVID complaints in persons with and without COVID-19. Sci Rep. 2023;13(1):6074. [FREE Full text] [doi: 10.1038/s41598-023-32636-y] [Medline: 37055494]
7. Mizrahi B, Sudry T, Flaks-Manov N, Yehezkelli Y, Kalkstein N, Akiva P, et al. Long covid outcomes at one year after mild SARS-CoV-2 infection: nationwide cohort study. BMJ. 2023;380:e072529. [FREE Full text] [doi: 10.1136/bmj-2022-072529] [Medline: 36631153]

8. Al-Aly Z, Bowe B, Xie Y. Long COVID after breakthrough SARS-CoV-2 infection. Nat Med. 2022;28(7):1461-1467. [FREE Full text] [doi: 10.1038/s41591-022-01840-0] [Medline: 35614233]
9. Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. Nature. 2021;594(7862):259-264. [doi: 10.1038/s41586-021-03553-9] [Medline: 33887749]
10. Xie Y, Choi T, Al-Aly Z. Association of treatment with nirmatrelvir and the risk of post-COVID-19 condition. JAMA Intern Med. 2023;183(6):554-564. [FREE Full text] [doi: 10.1001/jamainternmed.2023.0743] [Medline: 36951829]
11. Bowe B, Xie Y, Al-Aly Z. Acute and postacute sequelae associated with SARS-CoV-2 reinfection. Nat Med. 2022;28(11):2398-2405. [FREE Full text] [doi: 10.1038/s41591-022-02051-3] [Medline: 36357676]
12. Cohen K, Ren S, Heath K, Dasmariñas MC, Jubilo KG, Guo Y, et al. Risk of persistent and new clinical sequelae among adults aged 65 years and older during the post-acute phase of SARS-CoV-2 infection: retrospective cohort study. BMJ. 2022;376:e068414. [FREE Full text] [doi: 10.1136/bmj-2021-068414] [Medline: 35140117]
13. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health. 2016;37:61-81. [FREE Full text] [doi: 10.1146/annurev-publhealth-032315-021353] [Medline: 26667605]
14. Pfaff ER, Madlock-Brown C, Baratta JM, Bhatia A, Davis H, Girvin A, N3C Consortium, et al. RECOVER Consortium. Coding long COVID: characterizing a new disease through an ICD-10 lens. BMC Med. 2023;21(1):58. [FREE Full text] [doi: 10.1186/s12916-023-02737-6] [Medline: 36793086]
15. OMOP common data model. OHDSI. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ [accessed 2025-01-16]
16. Common Data Model (CDM) Specification, Version 6.1. PCORnet. URL: https://pcornet.org/wp-content/uploads/2022/01/PCORnet-Common-Data-Model-v60-2020_10_221.pdf [accessed 2025-01-16]
17. Carton TW, Marsolo K, Block JP. PCORnet COVID-19 common data model design and results. NIH Health Care Systems Collaboratory. 2020. URL: https://tinyurl.com/4zkbbavm [accessed 2025-01-16]
18. Pfaff ER, Girvin AT, Gabriel DL, Kostka K, Morris M, Palchuk MB, et al. Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. J Am Med Inform Assoc. 2022;29(4):609-618.
19. Pfaff ER, Girvin A, Crosskey M, Gangireddy S, Master H, Wei WQ, et al. N3CRECOVER Consortia. De-black-boxing health AI: demonstrating reproducible machine learning computable phenotypes using the N3C-RECOVER Long COVID model in the All of Us data repository. J Am Med Inform Assoc. 2023;30(7):1305-1312. [FREE Full text] [doi: 10.1093/jamia/ocad077] [Medline: 37218289]
20. Bergquist T, Loomba J, Pfaff E, Xia F, Zhao Z, Zhu Y, Long COVID Computational Challenge Participants, et al. N3C Consortium. Crowd-sourced machine learning prediction of long COVID using data from the national COVID cohort collaborative. EBioMedicine. 2024;108:105333. [FREE Full text] [doi: 10.1016/j.ebiom.2024.105333] [Medline: 39321500]
21. Brown JS, Bastarache L, Weiner MG. Aggregating electronic health record data for COVID-19 research-caveat emptor. JAMA Netw Open. 2021;4(7):e2117175. [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.17175] [Medline: 34255055]
22. Leese P, Anand A, Girvin A, Manna A, Patel S, Yoo YJ, et al. Clinical encounter heterogeneity and methods for resolving in networked EHR data: a study from N3C and RECOVER programs. J Am Med Inform Assoc. 2023;30(6):1125-1136. [FREE Full text] [doi: 10.1093/jamia/ocad057] [Medline: 37087110]
23. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC). 2013;1(3):1035. [FREE Full text] [doi: 10.13063/2327-9214.1035] [Medline: 25848578]
24. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. BMC Med Res Methodol. 2021;21(1):234. [FREE Full text] [doi: 10.1186/s12874-021-01416-5] [Medline: 34706667]
25. Ledford H. How common is long COVID? Why studies give different answers. Nature. 2022;606(7916):852-853. [doi: 10.1038/d41586-022-01702-2] [Medline: 35725828]
26. Bastarache L, Brown JS, Cimino JJ, Dorr DA, Embi PJ, Payne PR, et al. Developing real-world evidence from real-world data: transforming raw data into analytical datasets. Learn Health Syst. 2022;6(1):e10293. [FREE Full text] [doi: 10.1002/lrh2.10293] [Medline: 35036557]
27. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med. 2024;7(1):6. [FREE Full text] [doi: 10.1038/s41746-023-00970-0] [Medline: 38200151]
28. Rao S, Lee GM, Razzaghi H, Lorman V, Mejias A, Pajor NM, et al. Clinical features and burden of postacute sequelae of SARS-CoV-2 infection in children and adolescents. JAMA Pediatr. 2022;176(10):1000-1009. [FREE Full text] [doi: 10.1001/jamapediatrics.2022.2800] [Medline: 35994282]
29. Ely EW, Brown LM, Fineberg HV, National Academies of Sciences, Engineering,Medicine Committee on Examining the Working Definition for Long Covid. Long covid defined. N Engl J Med. 2024;391(18):1746-1753. [doi: 10.1056/NEJMsb2408466] [Medline: 39083764]

30. Electronic health records–based phenotyping. NIH Health Care Systems Collaboratory. URL: https://rethinkingclinicaltrials.org/chapters/conduct/electronic-health-records-based-phenotyping/electronic-health-records-based-phenotyping-introduction/ [accessed 2025-01-16]
31. Mayor N, Meza-Torres B, Okusi C, Delanerolle G, Chapman M, Wang W, et al. Developing a long COVID phenotype for postacute COVID-19 in a national primary care sentinel cohort: observational retrospective database analysis. JMIR Public Health Surveill. 2022;8(8):e36989. [FREE Full text] [doi: 10.2196/36989] [Medline: 35861678]
32. Zang C, Zhang Y, Xu J, Bian J, Morozyuk D, Schenck EJ, et al. Data-driven analysis to understand long COVID using electronic health records from the RECOVER initiative. Nat Commun. 2023;14(1):1948. [FREE Full text] [doi: 10.1038/s41467-023-37653-z] [Medline: 37029117]
33. Fashina TA, Miller C, Paintsil E, Niccolai L, Brandt C, Oliveira C. Computable clinical phenotyping of postacute sequelae of COVID-19 in pediatrics using real-world data. J Pediatric Infect Dis Soc. 2023;12(2):113-116. [FREE Full text] [doi: 10.1093/jpids/piac132] [Medline: 36548966]
34. Binka M, Klaver B, Cua G, Wong AW, Fibke C, Velásquez García HA, et al. An elastic net regression model for identifying long COVID patients using health administrative data: a population-based study. Open Forum Infect Dis. 2022;9(12):ofac640. [FREE Full text] [doi: 10.1093/ofid/ofac640] [Medline: 36570972]
35. Pfaff ER, Girvin AT, Bennett TD, Bhatia A, Brooks IM, Deer RR, et al. N3C Consortium. Identifying who has long COVID in the USA: a machine learning approach using N3C data. Lancet Digit Health. 2022;4(7):e532-e541. [FREE Full text] [doi: 10.1016/S2589-7500(22)00048-6] [Medline: 35589549]
36. Lorman V, Rao S, Jhaveri R, Case A, Mejias A, Pajor NM, et al. Understanding pediatric long COVID using a tree-based scan statistic approach: an EHR-based cohort study from the RECOVER Program. JAMIA Open. 2023;6(1):ooad016. [FREE Full text] [doi: 10.1093/jamiaopen/ooad016] [Medline: 36926600]
37. Lorman V, Razzaghi H, Song X, Morse K, Utidjian L, Allen AJ, et al. A machine learning-based phenotype for long COVID in children: an EHR-based study from the RECOVER program. PLoS One. 2023;18(8):e0289774. [FREE Full text] [doi: 10.1371/journal.pone.0289774] [Medline: 37561683]
38. Hadley E, Yoo YJ, Patel S, Zhou A, Laraway B, Wong R, N3C and RECOVER consortia, et al. Insights from an N3C RECOVER EHR-based cohort study characterizing SARS-CoV-2 reinfections and Long COVID. Commun Med (Lond). 2024;4(1):129. [FREE Full text] [doi: 10.1038/s43856-024-00539-2] [Medline: 38992084]
39. Kenny G, McCann K, O'Brien C, Savinelli S, Tinago W, Yousif O, et al. All-Ireland Infectious Diseases (AIID) Cohort Study Group. Identification of distinct long COVID clinical phenotypes through cluster analysis of self-reported symptoms. Open Forum Infect Dis. 2022;9(4):ofac060. [FREE Full text] [doi: 10.1093/ofid/ofac060] [Medline: 35265728]
40. Reese JT, Blau H, Casiraghi E, Bergquist T, Loomba JJ, Callahan TJ, N3C Consortium, et al. RECOVER Consortium. Generalisable long COVID subtypes: findings from the NIH N3C and RECOVER programmes. EBioMedicine. 2023;87:104413. [FREE Full text] [doi: 10.1016/j.ebiom.2022.104413] [Medline: 36563487]
41. Zhang H, Zang C, Xu Z, Zhang Y, Xu J, Bian J, et al. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. Nat Med. 2023;29(1):226-235. [FREE Full text] [doi: 10.1038/s41591-022-02116-3] [Medline: 36456834]
42. Peluso MJ, Deeks SG. Mechanisms of long COVID and the path toward therapeutics. Cell. 2024;187(20):5500-5529. [FREE Full text] [doi: 10.1016/j.cell.2024.07.054] [Medline: 39326415]
43. Davis HE, McCorkell L, Vogel JM, Topol EJ. Long COVID: major findings, mechanisms and recommendations. Nat Rev Microbiol. 2023;21(3):133-146. [FREE Full text] [doi: 10.1038/s41579-022-00846-2] [Medline: 36639608]
44. Zhang HG, Honerlaw JP, Maripuri M, Samayamuthu MJ, Beaulieu-Jones BR, Baig HS, Consortium for Clinical Characterization of COVID-19 by EHR (4CE), et al. Potential pitfalls in the use of real-world data for studying long COVID. Nat Med. 2023;29(5):1040-1043. [FREE Full text] [doi: 10.1038/s41591-023-02274-y] [Medline: 37055567]
45. Deer RR, Rock MA, Vasilevsky N, Carmody L, Rando H, Anzalone AJ, et al. Characterizing long COVID: deep phenotype of a complex condition. EBioMedicine. 2021;74:103722. [FREE Full text] [doi: 10.1016/j.ebiom.2021.103722] [Medline: 34839263]
46. Brown CA, Londhe AA, He F, Cheng A, Ma J, Zhang J, et al. Development and validation of algorithms to identify COVID-19 patients using a US electronic health records database: a retrospective cohort study. Clin Epidemiol. 2022;14:699-709. [FREE Full text] [doi: 10.2147/CLEP.S355086] [Medline: 35633659]
47. Khera R, Mortazavi BJ, Sangha V, Warner F, Patrick Young H, Ross JS, et al. A multicenter evaluation of computable phenotyping approaches for SARS-CoV-2 infection and COVID-19 hospitalizations. NPJ Digit Med. 2022;5(1):27. [FREE Full text] [doi: 10.1038/s41746-022-00570-4] [Medline: 35260762]
48. Joung SY, Ebinger JE, Sun N, Liu Y, Wu M, Tang AB, et al. Awareness of SARS-CoV-2 omicron variant infection among adults with recent COVID-19 seropositivity. JAMA Netw Open. 2022;5(8):e2227241. [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.27241] [Medline: 35976645]
49. Sudre CH, Murray B, Varsavsky T, Graham MS, Penfold RS, Bowyer RC, et al. Attributes and predictors of long COVID. Nat Med. 2021;27(4):626-631. [FREE Full text] [doi: 10.1038/s41591-021-01292-y] [Medline: 33692530]
50. Yong SJ. Long COVID or post-COVID-19 syndrome: putative pathophysiology, risk factors, and treatments. Infect Dis (Lond). 2021;53(10):737-754. [FREE Full text] [doi: 10.1080/23744235.2021.1924397] [Medline: 34024217]

51. Subramanian A, Nirantharakumar K, Hughes S, Myles P, Williams T, Gokhale KM, et al. Symptoms and risk factors for long COVID in non-hospitalized adults. Nat Med. 2022;28(8):1706-1714. [FREE Full text] [doi: 10.1038/s41591-022-01909-w] [Medline: 35879616]
52. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: issues, methods, and directions for the future. Health Serv Res. 2010;45(5 Pt 2):1468-1488. [FREE Full text] [doi: 10.1111/j.1475-6773.2010.01142.x] [Medline: 21054367]
53. Brannock MD, Chew RF, Preiss AJ, Hadley EC, Redfield S, McMurry JA, et al. N3C and RECOVER consortia. Long COVID risk and pre-COVID vaccination in an EHR-based cohort study from the RECOVER program. Nat Commun. 2023;14(1):2914. [FREE Full text] [doi: 10.1038/s41467-023-38388-7] [Medline: 37217471]
54. DeBord DG, Carreón T, Lentz TJ, Middendorf PJ, Hoover MD, Schulte PA. Use of the "Exposome" in the practice of epidemiology: a primer on -omic technologies. Am J Epidemiol. 2016;184(4):302-314. [FREE Full text] [doi: 10.1093/aje/kwv325] [Medline: 27519539]
55. Polubriaginof FCG, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, et al. Challenges with quality of race and ethnicity data in observational databases. J Am Med Inform Assoc. 2019;26(8-9):730-736. [FREE Full text] [doi: 10.1093/jamia/ocz113] [Medline: 31365089]
56. Hill EL, Mehta HB, Sharma S, Mane K, Singh SK, Xie C, et al. Risk factors associated with post-acute sequelae of SARS-CoV-2: an N3C and NIH RECOVER study. BMC Public Health. 2023;23(1):2103. [FREE Full text] [doi: 10.1186/s12889-023-16916-w] [Medline: 37880596]
57. Treweek S, Forouhi NG, Narayan KMV, Khunti K. COVID-19 and ethnicity: who will research results apply to? Lancet. 2020;395(10242):1955-1957. [doi: 10.1016/S0140-6736(20)31380-5] [Medline: 32539937]
58. Julian McFarlane S, Occa A, Peng W, Awonuga O, Morgan SE. Community-based participatory research (CBPR) to enhance participation of racial/ethnic minorities in clinical trials: a 10-year systematic review. Health Commun. 2022;37(9):1075-1092. [doi: 10.1080/10410236.2021.1943978] [Medline: 34420460]
59. Richesson RL, Marsolo K, Douthit B, Staman K, Ho PM, Dailey D, et al. Enhancing the use of EHR systems for pragmatic embedded research: lessons from the NIH health care systems research collaboratory. J Am Med Inform Assoc. 2021;28(12):2626-2640. [FREE Full text] [doi: 10.1093/jamia/ocab202] [Medline: 34597383]
60. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183(8):758-764. [FREE Full text] [doi: 10.1093/aje/kwv254] [Medline: 26994063]
61. Burki T. The future of paxlovid for COVID-19. Lancet Respir Med. 2022;10(7):e68. [FREE Full text] [doi: 10.1016/S2213-2600(22)00192-8] [Medline: 35623373]
62. Adler NE, Stead WW. Patients in context--EHR capture of social and behavioral determinants of health. N Engl J Med. 2015;372(8):698-701. [doi: 10.1056/NEJMp1413945] [Medline: 25693009]
63. Clarke KE, Jones JM, Deng Y, Nycz E, Lee A, Iachan R, et al. Seroprevalence of infection-induced SARS-CoV-2 antibodies - United States, September 2021-February 2022. MMWR Morb Mortal Wkly Rep. 2022;71(17):606-608. [FREE Full text] [doi: 10.15585/mmwr.mm7117e3] [Medline: 35482574]
64. L Mandel H, Colleen G, Abedian S, Ammar N, Charles Bailey L, Bennett TD, et al. Risk of post-acute sequelae of SARS-CoV-2 infection associated with pre-coronavirus disease obstructive sleep apnea diagnoses: an electronic health record-based analysis from the RECOVER initiative. Sleep. 2023;46(9):zsad126. [FREE Full text] [doi: 10.1093/sleep/zsad126] [Medline: 37166330]
65. Goralnick E, Kaufmann C, Gawande AA. Mass-vaccination sites - an essential innovation to curb the covid-19 pandemic. N Engl J Med. 2021;384(18):e67. [doi: 10.1056/NEJMp2102535] [Medline: 33691058]
66. Hill E, Mehta H, Sharma S, Mane K, Singh SK, Xie C, N3C Consortium, et al. RECOVER Consortium. Risk factors associated with post-acute sequelae of SARS-CoV-2: an N3C and NIH RECOVER study. BMC Public Health. Oct 25, 2023;23(1):2103. [FREE Full text] [doi: 10.1186/s12889-023-16916-w] [Medline: 37880596]
67. Forrest CB, Burrows E, Mejias A, Razzaghi H, Christakis D, Jhaveri R, et al. Pediatrics. 2022;149(4):e2021055765. [doi: 10.1542/peds.2021-055765] [Medline: 35322270]
68. Tannous J, Pan AP, Potter T, Bako AT, Dlouhy K, Drews A, et al. Real-world effectiveness of COVID-19 vaccines and anti-SARS-CoV-2 monoclonal antibodies against postacute sequelae of SARS-CoV-2: analysis of a COVID-19 observational registry for a diverse US metropolitan population. BMJ Open. 2023;13(4):e067611. [FREE Full text] [doi: 10.1136/bmjopen-2022-067611] [Medline: 37019490]
69. RECOVER: researching COVID to enhance recovery. RECOVER. URL: https://recovercovid.org/ [accessed 2015-01-26]

Abbreviations

CDC: Centers for Disease Control and Prevention
CSC: clinical science core
CRN: clinical research network
EHR: electronic health record
GIS: Geographic Information Systems

ICD-10: International Statistical Classification of Diseases, Tenth Revision
ME/CFS: myalgic encephalomyelitis/chronic fatigue syndrome
ML: machine learning
N3C: National COVID Cohort Collaborative
NIH: National Institutes of Health
OMOP: Observational Medical Outcomes Partnership
OSA: obstructive sleep apnea
PCORI: Patient-Centered Outcomes Research Institute
PCORNet: National Patient-Centered Clinical Research Network
POTS: postural orthostatic tachycardia syndrome
RECOVER: REsearching COVID to Enhance Recovery
VA: Department of Veterans Affairs

Edited by A Mavragani; submitted 02.05.24; peer-reviewed by Y Zhang, S Zeng; comments to author 12.09.24; revised version received 31.10.24; accepted 20.11.24; published 05.03.25

Please cite as:
Mandel HL, Shah SN, Bailey LC, Carton T, Chen Y, Esquenazi-Karonika S, Haendel M, Hornig M, Kaushal R, Oliveira CR, Perlowski AA, Pfaff E, Rao S, Razzaghi H, Seibert E, Thomas GL, Weiner MG, Thorpe LE, Divers J, RECOVER EHR Cohort
Opportunities and Challenges in Using Electronic Health Record Systems to Study Postacute Sequelae of SARS-CoV-2 Infection: Insights From the NIH RECOVER Initiative
J Med Internet Res 2025;27:e59217
URL: https://www.jmir.org/2025/1/e59217
doi: 10.2196/59217
PMID: 40053748