Identifying Long COVID Definitions, Predictors, and Risk Factors using Electronic Health Records: A Scoping Review

Rayanne Anderson Luke  
George Mason University

George Shaw, Jr.  
University of North Carolina at Charlotte

Geetha S. Clarke  
Stanford University

Abolfazl Mollalo (mollalo@musc.edu)  
Medical University of South Carolina  https://orcid.org/0000-0001-5092-0698

Research Article

Keywords: Electronic Health Records, Long COVID, Phenotypes, Post-acute COVID-19

Posted Date: December 4th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3689967/v1

License: ©  This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Objective

Long COVID, or post-COVID condition, is characterized by a range of physical and psychological symptoms and complications that persist beyond the acute phase of the coronavirus disease of 2019 (COVID-19). However, this condition still lacks a clear definition. This scoping review explores the potential of electronic health records (EHR)-based studies to characterize long COVID.

Methods

We screened all peer-reviewed publications in the English language from PubMed/MEDLINE, Scopus, and Web of Science databases until September 14, 2023. We identified studies that defined or characterized long COVID based on EHR data, regardless of geography or study design. We synthesized these articles based on their definitions, symptoms, and predictive factors or phenotypes to identify common features and analytical methods.

Results

We identified only 20 studies meeting the inclusion criteria, with a significant majority (n = 17, 85%) conducted in the United States. Respiratory conditions were significant in all studies, followed by poor well-being features (n = 17, 85%) and cardiovascular conditions (n = 14, 70%). Some articles (n = 8, 40%) used a long COVID-specific marker to define the study population, relying mainly on International Classification of Diseases, Tenth Revision (ICD-10) codes and clinical visits for post-COVID conditions. Among studies exploring plausible long COVID (n = 12, 60%), reverse transcription-polymerase chain reaction and antigen tests were the most common identification methods. The time delay for EHR data extraction post-test varied, ranging from four weeks to more than three months; however, most studies considering plausible long COVID used a waiting period of 28 to 31 days.

Conclusion

Our findings suggest a limited global utilization of EHR-derived data in defining or characterizing long COVID, with 60% of these studies incorporating a validation step. Future meta-analyses are essential to assess the homogeneity of results across different studies.

1. Introduction

Post-COVID conditions, known as long COVID, refer to symptoms that manifest about four or more weeks after the initial infection by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes the coronavirus disease of 2019 (COVID-19) [1, 2]. Patients with long COVID experience many symptoms and conditions affecting various organ systems, including the respiratory, circulatory, nervous, endocrine, and digestive systems [3]. Moreover, chronic medical conditions and certain risk factors such as high blood pressure, diabetes, and obesity can exacerbate symptoms [4]. The increase in medical spending in the United States (US) as a result of long COVID is estimated at $528 billion total (about $1,570 per person), about 2.5% of the total gross domestic product in 2019 [5]. A historical cohort study in Israel estimated the excess cost of long COVID on healthcare utilization as an additional 7.6% of control patient care [6]. Another study estimated productivity-based losses to the German economy at $3.7 billion [7]. Due to the
debilitating conditions of long COVID, functional impairment, low work productivity, and long-term health complications are anticipated [8, 9].

At present, little is known about the pathophysiology of these multisystem complications. While several studies have attempted to characterize long COVID [10–13], comparing the findings proves difficult due to the heterogeneity of symptoms and study period. Further, the accepted definition of long COVID keeps evolving [14]. The largest initiative in the US to define long COVID has been through the National Institute of Health (NIH)’s Researching COVID to Enhance Recovery (RECOVER) program. The program identified the 12 most prevalent symptom patterns among those with long COVID, including post-exercise malaise, fatigue, brain fog, dizziness, gastrointestinal symptoms, heart palpitations, changes in sexual desire or capacity, loss of taste or smell, thirst, chronic cough, chest pain, and abnormal movements [15]. As of November 2023, about 450 studies listed on ClinicalTrials.gov are investigating long COVID, including the response to antivirals, lithium therapy, and nitrate supplements, to discover possible treatments [16]. Several studies identified predictors or risk factors of long COVID [17–19] but used survey-based approaches that are prone to question heterogeneity and voluntary response bias. Identified examples included increasing age and body mass index (BMI), female sex, frailty, experiencing more than five symptoms or a hospital or emergency room visit during acute COVID-19 illness, and comorbidities such as asthma and heart disease [17–19].

Electronic health records (EHR) can provide a reliable, cost-effective, and comprehensive overview of the medical histories of a large population of patients, including previous infections and pre-existing conditions, thus enabling the tracking of symptoms and conditions [20]. Within EHR systems, a valuable resource for obtaining disease etiology is the International Classification of Diseases, Tenth Revision (ICD-10), used to code and classify all symptoms, procedures, and diagnoses [21]. Unexplainable symptoms of COVID-19 were classified as B94.8 as a placeholder to signify long COVID until September 30, 2021; the code was subsequently changed to U09 [22].

While previous review studies attempted to define long COVID [23–26], these efforts are constrained by their heterogeneous designs and lack of specific diagnostic definitions. For example, a previous meta-analysis aimed at characterizing long COVID included 39 studies of diverse study designs, such as cohort, cross-sectional, and case-control [23]. Among these studies, the majority (n = 34) had either a moderate or high risk of bias. Likewise, Kelly et al. [27] identified a wide spectrum of symptoms and significant heterogeneity across the studies, noting that the population was limited to hospitalized patients. An earlier review by Iqbal et al. [28] incorporated 35 articles until March 2021 and highlighted limitations pertaining to various study designs, a limited number of countries, and questionnaire-based cross-sectional studies that fail to capture the evolution of symptoms over time. However, to our knowledge, no review article has comprehensively explored the potential of EHR-based studies to characterize long COVID. Thus, this scoping review aims to collate the studies that defined long COVID based on phenotypes or identified predictors or risk factors derived from EHR data. We also identified the analytical methods and summarized the common significant phenotypes, predictors, or risk factors.

2. Methods

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to conduct this scoping review.

2.1. Data Source

We used three search engines, PubMed/MEDLINE, Scopus, and Web of Science, to identify peer-reviewed articles until September 14, 2023, regardless of geography or study design.

2.2. Search Strategy
The initial screening of abstracts and titles included all articles that contained our key terms pertaining to long COVID, and those that contained EHR or electronic medical records (EMR) (Table 1). Studies were considered eligible if they were written in English, were peer-reviewed, and defined or characterized long COVID phenotypes or its predictors or risk factors. Review articles, pre-prints, case reports, editorials, abstracts, and articles published in early 2020 and before were excluded. Notably, studies that used non-EHR-based features of long COVID were excluded.

Table 1
Search terms input into PubMed, Scopus, and Web of Science.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Key terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long COVID</td>
<td>&quot;long COVID*&quot; OR &quot;long-term COVID*&quot; OR &quot;post-acute sequelae&quot; OR &quot;late-stage COVID*&quot; OR &quot;SARS-CoV-2 post-recovery&quot; OR &quot;PASC&quot; OR &quot;long-haul COVID*&quot; OR &quot;Chronic COVID*&quot; OR &quot;persistent COVID*&quot; OR &quot;prolonged COVID*&quot; OR &quot;extended COVID*&quot; OR &quot;post-recovery COVID*&quot; OR &quot;Aftermath COVID*&quot; OR &quot;survivorship COVID*&quot; OR &quot;late effects COVID*&quot; OR &quot;long-term effects of COVID&quot; OR &quot;post-acute COVID-19&quot; OR &quot;post-acute sequelae of SARS-CoV-2 (PASC)&quot; OR &quot;ICD-10-CM&quot; OR &quot;ICD-10&quot;</td>
</tr>
<tr>
<td>Electronic health records</td>
<td>&quot;Electronic Medical Record*&quot; OR &quot;Electronic Health Record*&quot; OR &quot;Electronic Patient Record*&quot; OR &quot;EHR&quot; OR &quot;EMR&quot; OR &quot;N3C&quot; OR &quot;All of US&quot;</td>
</tr>
</tbody>
</table>

2.3. Study Selection

We imported all the retrieved abstracts and titles into the Covidence online tool (https://www.covidence.org/), where duplicates are excluded. Two reviewers (AM and GSC) independently screened abstracts and titles and excluded the articles not about long COVID. Next, two different reviewers (RAL and GSJ) screened the full texts of the selected articles. In this step, articles were excluded if they did not have a clear definition of long COVID or did not identify risk factors, phenotypes, or predictors. Conflicts were resolved through follow-up discussions to reach a consensus on the final selection of articles. We also conducted a reference check of the selected articles for possible inclusion. Figure 1 shows the process of study selection.

2.4. Data Extraction

We extracted information from the selected articles, including author names, study dates, sample size, types of data collected (qualitative and quantitative), study population inclusion criteria, identified long COVID features, methods used to identify such features, methods used to validate results, and study limitations.

2.5. Narrative Synthesis

A summary of long COVID labels assigned to samples in each study was created, and data was categorized based on clinical phenotypes and databases used in each of these papers. We synthesized various methods used to identify the important characteristics of long COVID phenotypes, predictors, and risk factors.

3. Results

Initially, we identified 543 articles from PubMed, Web of Science, and Scopus databases. Of these, 327 studies were duplicates and removed, resulting in 216 unique studies for the initial screening. After the abstract and title screening of those articles, another 113 studies were excluded. We conducted the full-text review of the remaining 103 studies and excluded 93 studies due to a lack of a clear definition of long COVID and/or missing identification of predictors or risk factors. We identified ten additional studies via a reference check of the articles. Therefore, the final selection comprised 20 studies. Table 2 provides a summary of the selected studies.

The total sample size of long COVID study and control populations ranged from 11,209 to 5,213,885. The Veterans Affairs Electronic Healthcare database (n = 5,808,018), the Consortium for the Clinical Characterization of COVID-19 by
EHR (4CE) (n = 5,434,528), and INSIGHT Network, New York City Health System (n = 5,346,357) provided the largest databases of potential patients. The size of the 4CE database of positive COVID-19 patients varied among each study as new data were added monthly. Additionally, some datasets used inpatient and outpatient records to identify positive cases of COVID. The geographic distribution of articles revealed that 85% of the selected studies utilized datasets composed of US patients. The remaining studies were carried out on patient data from France, Germany, Italy, and Singapore. The authors used qualitative (20%), quantitative (20%), and a mixture (30%) of qualitative and quantitative data. The remaining did not provide explicit evidence, or the data was unclear.

Figure 2 shows the methods for identifying candidate patient labels, classified as long COVID (n = 8/20) or plausibly long COVID (n = 12/20). Among articles that used a long COVID-specific marker (Fig. 2a), most (6/8, 75%) chose an ICD-10 code for post COVID-19 condition, unspecified, or sequelae of other specified infectious diseases. Among the studies that used acute COVID-19 illness to define plausibly long COVID patients (n = 12/20), reverse transcription-polymerase chain reaction (RT-PCR) and antigen tests were most common (Fig. 2b). These twelve studies extracted features to study from the EHR after a waiting period following a recorded acute COVID-19 positive test. The waiting period shown in Fig. 2b varied from 28 days to more than three months [29, 30], but most used a one-month delay.
<table>
<thead>
<tr>
<th>Study*</th>
<th>Year</th>
<th>Study Design</th>
<th>Total Sample Size</th>
<th>Total Participants (n = study population)</th>
<th>EHR Source</th>
<th>Main Methods</th>
<th>Method of Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al-Aly et al. [31]</td>
<td>2021</td>
<td>Observational Study</td>
<td>5,213,885</td>
<td>73,435</td>
<td>US Department of Veterans Affairs electronic healthcare databases</td>
<td>Cox regression</td>
<td>N/A</td>
</tr>
<tr>
<td>Baskett et al. [32]</td>
<td>2022</td>
<td>Retrospective Cohort Study</td>
<td>N/A</td>
<td>17,487</td>
<td>Cerner Real-World data set</td>
<td>Logistic regression; Propensity score matching</td>
<td>N/A</td>
</tr>
<tr>
<td>Dagliati et al. [29]</td>
<td>2023</td>
<td>Retrospective Cohort Study</td>
<td>30,422</td>
<td>12,424</td>
<td>4CE</td>
<td>Clustering; MLHO</td>
<td>Clinical expert review</td>
</tr>
<tr>
<td>Estiri et al. [30]</td>
<td>2021</td>
<td>Retrospective Cohort Study</td>
<td>96,025</td>
<td>22,475</td>
<td>Mass General Brigham Hospital</td>
<td>MLHO; Multivariate time series analysis</td>
<td>Clinical expert review</td>
</tr>
<tr>
<td>Fritsche et al. [33]</td>
<td>2023</td>
<td>Case-Control</td>
<td>63,675</td>
<td>1,724</td>
<td>Michigan Medicine EHR data</td>
<td>Logistic regression</td>
<td>Sensitivity analysis; Logistic regression; AAUC</td>
</tr>
<tr>
<td>Haupert et al. [34]</td>
<td>2022</td>
<td>Case-Crossover Design</td>
<td>204,597</td>
<td>44,198</td>
<td>Michigan Medicine Health System</td>
<td>Logistic regression</td>
<td>N/A</td>
</tr>
<tr>
<td>Jiang et al. [35]</td>
<td>2022</td>
<td>Retrospective Cohort Study</td>
<td>85,196</td>
<td>28,558</td>
<td>N3C EHR repository</td>
<td>Deep neural networks; PCA</td>
<td>AUC; F1 score</td>
</tr>
<tr>
<td>Kessler et al. [36]</td>
<td>2023</td>
<td>Retrospective Exploratory Study</td>
<td>272,588</td>
<td>5,440</td>
<td>IQVIA Disease Analyzer Database</td>
<td>Light gradient boosting (decision tree)</td>
<td>AUC; Precision; Recall; Specificity; F2 Score</td>
</tr>
<tr>
<td>Khullar et al. [37]</td>
<td>2023</td>
<td>Retrospective Cohort Study</td>
<td>310,220</td>
<td>62,339</td>
<td>INSIGHT Network; New York City Health Systems</td>
<td>Logistic regression</td>
<td>Sensitivity analysis</td>
</tr>
<tr>
<td>Lorman et al. [38]</td>
<td>2023</td>
<td>Observational Study</td>
<td>14,399</td>
<td>1,309</td>
<td>RECOVER PEDSNet EHR</td>
<td>Propensity score matching; Decision trees</td>
<td>Sensitivity analysis</td>
</tr>
<tr>
<td>Study*</td>
<td>Year</td>
<td>Study Design</td>
<td>Total Sample Size</td>
<td>Total Participants (n = study population)</td>
<td>EHR Source</td>
<td>Main Methods</td>
<td>Method of Validation</td>
</tr>
<tr>
<td>-----------------</td>
<td>------</td>
<td>-----------------------------</td>
<td>-------------------</td>
<td>------------------------------------------</td>
<td>-----------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Nasir et al. [39]</td>
<td>2023</td>
<td>Observational Study</td>
<td>11,209</td>
<td>4,091</td>
<td>Health Choice Network</td>
<td>Bayesian structural time series modeling</td>
<td>N/A</td>
</tr>
<tr>
<td>Pfaff et al. [40]</td>
<td>2023</td>
<td>Retrospective Cohort Study</td>
<td>36,880</td>
<td>33,782</td>
<td>N3C</td>
<td>Clustering; network analysis</td>
<td>N/A</td>
</tr>
<tr>
<td>Pfaff et al. [22]</td>
<td>2022</td>
<td>Retrospective Cohort Study</td>
<td>1,793,604</td>
<td>73,972</td>
<td>N3C</td>
<td>Extreme gradient boosting (decision tree)</td>
<td>Cross-validation; AUC; Precision; Recall; F-score</td>
</tr>
<tr>
<td>Rao et al. [41]</td>
<td>2022</td>
<td>Exploratory, Retrospective Cohort Study</td>
<td>659,286</td>
<td>59,893</td>
<td>PEDSnet</td>
<td>Cox regression; logistic regression</td>
<td>N/A</td>
</tr>
<tr>
<td>Reese et al. [42]</td>
<td>2023</td>
<td>Retrospective Cohort Study</td>
<td>5,434,528</td>
<td>20,532</td>
<td>N3C</td>
<td>NLP; k-means clustering</td>
<td>N/A</td>
</tr>
<tr>
<td>Sengupta et al. [43]</td>
<td>2022</td>
<td>Retrospective Cohort Study</td>
<td>49,950</td>
<td>7,511</td>
<td>N3C</td>
<td>Convolutional and LSTM neural networks</td>
<td>AUC</td>
</tr>
<tr>
<td>Wang et al. [44]</td>
<td>2022</td>
<td>Observational Study</td>
<td>51,485</td>
<td>26,117</td>
<td>Mass General Brigham Hospital</td>
<td>Rule-based NLP</td>
<td>Precision</td>
</tr>
<tr>
<td>Zang et al. [45]</td>
<td>2023</td>
<td>Observational Study</td>
<td>361,401 (INSIGHT)</td>
<td>35,275 (INSIGHT)</td>
<td>INSIGHT CRN and OneFlorida + CRN</td>
<td>Propensity score matching</td>
<td>Sensitivity analysis</td>
</tr>
<tr>
<td>HG Zhang et al. [46]</td>
<td>2022</td>
<td>Retrospective Cohort Study</td>
<td>2,745,130</td>
<td>414,602</td>
<td>4CE</td>
<td>Random-effects meta-analysis</td>
<td>N/A</td>
</tr>
<tr>
<td>H Zhang et al. [47]</td>
<td>2022</td>
<td>Retrospective Cohort Study</td>
<td>34,605</td>
<td>20,881</td>
<td>INSIGHT CRN and OneFlorida + CRN</td>
<td>Clustering; PFA topic modeling</td>
<td>Topic Coherence; Sensitivity analysis</td>
</tr>
</tbody>
</table>

*Abbreviations: 4CE: Consortium of Clinical Characterization of COVID-19 by EHR; AUC: area under the covariate-adjusted receiver operating characteristic curve; AUC: area under the receiver operating characteristic curve; ANOVA: analysis of variance; CRN: clinical research network MLHO: Machine Learns Health Outcomes; PEDSnet: a National Pediatric Learning Health System; N3C: National COVID Cohort Collaborative; NLP: natural language processing; LSTM: long short-term memory; PCA: principal component analysis; PFA: Poisson factor analysis.

Figure 3 depicts the candidates used to define long COVID or identify predictors or risk factors. All studies used diagnostic categories as candidates. Many studies organized their diagnostic category candidates by grouping ICD-10 codes into diagnostic clusters using a version of Clinical Classifications Software Refined. In contrast, some used the R
package PheWAS to map ICD-10 codes to unique PheCodes. Notably, seven (35%) studies included medications, and six (30%) studies used a combination of social determinants of health (SDOH) and patient demographics. Only one study [35] used quantitative data based on biophysical metrics recorded during acute COVID-19 hospitalizations.

The main data analysis methods used by the studies are presented in Fig. 4. In this figure, we depicted the methods used by at least two studies. See Table 2 for methods for each study. Due to the relatively large sample sizes, most studies used data science techniques. Ten (50%) articles discussed the implementation of machine learning for feature down-selection or identification of significant features. Various clustering methods were utilized, including principal component analysis, the Louvain algorithm, probabilistic topic modeling, and hierarchical agglomerative clustering. Moreover, several studies implemented decision trees, two of which used the extreme gradient boosting algorithm.

All articles incorporated statistical tests, but approaches varied. Nearly all articles (n = 17, 85%) reported results using hazard ratios, risk ratios, odds ratios, or p-values. Hazard ratios were often found using Cox survival models, and reported p-values often underwent the Bonferroni correction. Some studies reduced their initial pool of candidates to a final set based on feature importance criteria such as hazard ratios greater than 1 or p-values less than 0.05. In contrast, others merely discussed the top 20 or 30 features.

We categorized the phenotypes, predictors, and risk factors of long COVID identified by the articles (Fig. 5). As some articles used hundreds of candidates in their analyses, Fig. 5 only shows the categories of most significant findings. The studies varied significantly in how they clustered and summarized their results; we grouped these by broad categories to facilitate comparison. All articles found respiratory conditions significant, nearly all studies (n = 17, 85%) reported poor well-being features as significant, and most (n = 14, 70%) included cardiovascular conditions. Examples for these top three categories were: respiratory conditions (e.g., respiratory failure, asthma), poor general well-being (e.g., fatigue, pain), and cardiovascular conditions and diseases (e.g., chest pain, coronary disease). Fifty-five percent of the definitions, predictors, and risk factors used to understand long COVID were qualitative; 15% of the articles used a quantitative approach, 5% were qualitative and quantitative, and 25% had either an incomplete or nonexistent description.

Twelve of the studies (60%) included a validation step to test their purported long COVID phenotype, predictors, or risk factors on unlabeled data. One-fourth used their entire suite of candidate features in a long COVID prediction task and measured their success with the area under the receiver operating characteristic curve (AUC) or F scores. A quarter of the studies (25%) used sensitivity analysis to test the robustness of their results. Wang et al. [44] used natural language processing (NLP) to validate their reduced symptom lexicon on a test dataset and achieved an average precision and estimated recall of 0.94 and 0.84. Dangliati et al. and Estiri et al. [29, 30] employed clinical experts to review the phenotypes identified by their method.

4. Discussion

This is the first scoping review for identifying long COVID phenotypes, predictors, and risk factors based on EHR data. We found 20 articles that met the eligibility criteria. The articles were classified based on how long COVID was defined, methodologies, and the identification of significant risk factors or phenotypes. ICD-10 codes (U09.9 or B 94.8) were the most common markers. A large majority of the studies reported poor general well-being and respiratory and cardiovascular conditions features as significant.

The studies collectively illuminate the heterogeneous nature of long COVID, revealing that it does not manifest as a uniform condition. Instead, it presents diverse phenotypic clusters, which include respiratory, neuropsychiatric, cardiovascular, and pain and fatigue subtypes [42]. These subtypes are characterized by distinct clinical features, patient demographics, and associations with various organ systems. Ten of the twelve symptom patterns identified by the NIH
RECOVER initiative were considered as features by the studies and fell under broad, significant categories (Fig. 5). Both thirst and changes in sexual desire were absent as individual features but may fall under broad symptom characterizations [42]. Phenotypic clustering utilizing machine learning methods may provide a unique approach to examining phenotype commonalities and drastic differences, and may enhance our understanding of long COVID’s heterogeneity. While several of the phenotypes found by the studies are common in understanding long COVID, latent phenotypes like substance use disorders (e.g., opioid use) and genital organs present unique ways to define the disease.

Beyond the symptoms of long COVID, some studies reported predictors and risk factors for developing the condition. We noted that SDOH and demographic information cannot plausibly be used to form a clinical definition of long COVID but may indicate the likelihood of developing the condition. Such studies included patients who recovered quickly from their acute infection in their test population, therefore diluting the conclusions that can be drawn from observed symptoms, diagnoses, and medicines.

Surprisingly, our research found limited global use of EHR data in defining or characterizing long COVID, with 85% of the included studies utilizing datasets composed of US patients. These geographical disparities in research distribution could mainly be related to privacy regulations governing access to EHR data across different countries. The lack of diverse EHR-based studies may hinder our understanding of long COVID’s complex nature and the generalizability of findings beyond the studied regions. Thus, it is recommended that future international studies harness the full potential of EHR data for long COVID research, enabling the development of more effective interventions.

The studies identified a broad range of symptoms, complications, and clinical conditions significantly associated with long COVID. However, only one study [39] examined causality, using Bayesian structural time series models. This complexity underscores the need for a more comprehensive understanding of clinical conditions to elucidate the causal relationships. Moreover, while the studies primarily relied on ICD-10 codes to define long COVID, only one study leveraged the potential of NLP for extracting information from unstructured clinical notes in EHR data. Combined with longitudinal EHR data, this approach can enhance our understanding of patterns, progression, treatment, and management of long COVID.

While EHR, EMR, and electronic patient records (EPR) systems are often used interchangeably, they differ in scope and usage. EHRs were the most frequently encountered systems in this review. They serve as digital repositories of a patient’s complete medical history from all healthcare providers and are intended for sharing with other healthcare entities, practices, and hospitals. EMRs focus on a single practice or hospital. Some EMR systems offer integrated care, possibly within larger hospital corporations. Lastly, EPR usage is primarily limited to Europe, which was not used in the articles. The localized usage of the term EPR may indicate the use of other terminologies globally that we have not accounted for in our literature search strategy, as our search was confined to specific health information systems. For future studies, it would be beneficial to include broader research terms, such as data hubs, data lakes, registries, and repositories, to encompass a more comprehensive scope of relevant literature.

The included studies shared some common limitations. There is no widely accepted definition for long COVID to date, so the disease was defined based on various arbitrary time intervals. Additionally, the symptoms and conditions of long COVID were examined mostly in the initial months. Most studies failed to differentiate between incident and prevalent symptoms, contributing to ambiguity in characterizing long COVID. To address these challenges, a standardized research protocol that facilitates capturing systematic tracking of symptoms seems imperative. Leveraging longitudinal EHR-derived data can provide a more comprehensive understanding of the emerging symptoms and monitor longer-term trends and effects.

Nearly all studies acknowledged data-related limitations. The use of diagnosis codes is a limitation because they may not capture all signs, symptoms, or laboratory results found in clinical notes. This reliance on diagnostic codes could...
result in missing information and biases. Additionally, the intensity of the pandemic and misinformation might introduce confirmatory bias between healthcare providers and patients. Some articles [34, 45] acknowledged the need for replication studies in other cohorts. Several articles excluded hospitalized COVID-19 patients, which might not reflect the complete spectrum of long COVID. This limitation highlights the importance of including more hospitalized patients in future studies.

Temporal bias is an issue because the choice of time windows for analysis varies among studies. Some studies expressed concerns about not accounting for within-person time-varying confounders, such as changes in health-seeking behavior during the pandemic. The studies’ requirement of pre- and post-COVID-19 visits is acknowledged as potentially biased toward patients with more complex health histories. The cohort case window ratio is a subjective parameter, indicating variability in the study design. Additionally, some studies mentioned potential timing biases related to developing and implementing specific diagnosis codes. Some touched on the issue of data representation concerning various population groups and the inability to generalize the data. The absence of medication information used for COVID-19 therapy, particularly in severe cases, is noted as a potential limitation, as is the lack of data on viral variants for individual patients.

Our study has several limitations. First, the exclusion of non-English articles may introduce language bias. Due to the evolving nature of the topic, the omission of pre-prints and conference articles could result in the loss of crucial information. Additionally, it is possible that some articles were missed during the screening process, although we mitigated this by conducting reference checks and backward searches. Another recurring limitation was the lack of comprehensive insight into the underlying mechanisms that lead to long COVID. Furthermore, they did not explore the indirect effects of long COVID, such as the social, economic, and behavioral changes that may arise due to the condition. This knowledge gap poses a significant challenge in fully understanding and addressing long COVID.

In conclusion, this scoping review has provided a comprehensive overview of state-of-the-art research on long COVID, utilizing EHR data as the primary source for defining and characterizing this condition. The findings suggest that while a consensus on the definition of long COVID remains elusive, ICD-10 codes are commonly used for identification, and poor general well-being, respiratory conditions, and cardiovascular conditions are consistently associated with long COVID. It was observed that the use of EHR data for characterizing long COVID is primarily concentrated in the US, highlighting the need for more international studies. Moreover, while data science techniques are widely employed, the lack of validation and causality assessments is evident, highlighting the need for more robust methodologies. The complex nature of long COVID, encompassing various symptoms and clinical conditions, underscores the need for more in-depth studies, including those leveraging longitudinal EHR data. Despite the gaps, this review serves as a foundation for future research efforts aimed at harnessing the potential of EHR data to better understand the epidemiology of long COVID.

**Declarations**

**Competing interests:**

None.

**Funding:**

None.

**Author contributions:**
Study concept and design: AM; Title and abstract screening: GSC and AM; Full-text review: RAL and GSJ; Data extraction: All authors; Writing and editing and final approval: All authors.

References


Figures
Figure 1

PRISMA flowchart of EHR-based characterization of long COVID.
Markers used for candidate patient labels: (a) long COVID and (b) plausibly long COVID. Plausible long COVID markers were paired with a time duration post-test.

Candidates for the definition, predictors, or risk factors of long COVID.
Figure 4

Main data analysis method used. Methods in the Other category: network analysis and random-effects meta-analysis.

Figure 5
Significant phenotypes, predictors, and risk factors of long COVID.