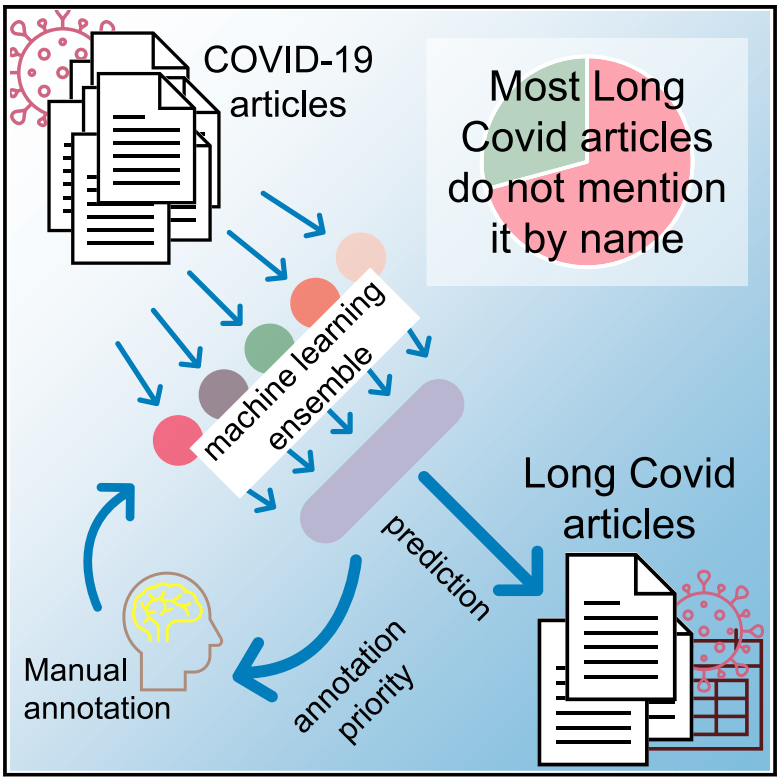
Descriptor

Patterns

Comprehensively identifying Long Covid articles with human-in-the-loop machine learning

Graphical abstract



Highlights

- We classify COVID-19 articles for relevance to Long Covid, a novel condition
- Most Long Covid articles do not mention it by name, complicating identification
- We ensemble differing data views for robust prediction and to direct human annotation
- We created the Long Covid Collection, ~9,000 articles, available in LitCovid

Authors

Robert Leaman, Rezarta Islamaj, Alexis Allot, Qingyu Chen, W. John Wilbur, Zhiyong Lu

Correspondence

zhiyong.lu@nih.gov

In brief

A significant percentage of COVID-19 survivors experience Long Covid: ongoing multisystemic symptoms that often affect daily living. Querying for relevant scientific articles is difficult before consensus builds for standardized terminology; we therefore identified Long Covid articles using an iterative human-in-the-loop machine learning approach. Analysis of the ~9,000 articles in the Long Covid Collection shows that most do not mention it by name and emphasizes that Long Covid is associated with disorders in a wide variety of body systems.

Leaman et al., 2023, Patterns *4*, 100659
January 13, 2023
https://doi.org/10.1016/j.patter.2022.100659

Patterns

CellPress
OPEN ACCESS

Descriptor

Comprehensively identifying Long Covid articles with human-in-the-loop machine learning

Robert Leaman,[1] Rezarta Islamaj,[1] Alexis Allot,[1] Qingyu Chen,[1] W. John Wilbur,[1] and Zhiyong Lu[1,2,*]
[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health 8600 Rockville Pike, Bethesda, MD 20894, USA
[2]Lead contact
*Correspondence: zhiyong.lu@nih.gov
https://doi.org/10.1016/j.patter.2022.100659

THE BIGGER PICTURE Long Covid causes ongoing multisystemic symptoms in a substantial percentage of COVID-19 survivors and lacks specific treatments. Locating articles that refer to novel entities such as Long Covid is generally challenging since keyword searches suffer from limited results and low accuracy without broadly supported terminology. We developed an iterative human-in-the-loop framework to comprehensively identify articles relevant to Long Covid. Our framework integrates multiple classifiers with complementary views and varying accuracy into a single model that reliably predicts the relevance of each article to Long Covid and its priority for manual annotation. We show that most articles relevant to Long Covid do not name the condition and are missed by keyword search. We present and analyze a comprehensive collection of Long Covid articles in LitCovid, which we believe will help accelerate research into this pressing public health issue.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

A significant percentage of COVID-19 survivors experience ongoing multisystemic symptoms that often affect daily living, a condition known as Long Covid or post-acute-sequelae of SARS-CoV-2 infection. However, identifying scientific articles relevant to Long Covid is challenging since there is no standardized or consensus terminology. We developed an iterative human-in-the-loop machine learning framework combining data programming with active learning into a robust ensemble model, demonstrating higher specificity and considerably higher sensitivity than other methods. Analysis of the Long Covid Collection shows that (1) most Long Covid articles do not refer to Long Covid by any name, (2) when the condition is named, the name used most frequently in the literature is Long Covid, and (3) Long Covid is associated with disorders in a wide variety of body systems. The Long Covid Collection is updated weekly and is searchable online at the LitCovid portal: https://www.ncbi.nlm.nih.gov/research/coronavirus/docsum?filters=e_condition.LongCovid.

INTRODUCTION

Literature collections such as LitCovid provide a critical resource as scientific understanding expands, serving as a centralized access point for reliable and comprehensive information on COVID-19.[1] LitCovid initially launched in February 2020, providing a set of eight topics, such as prevention and diagnosis, to improve information accessibility.[2] As our understanding of the effects of COVID-19 continues to evolve, however, updates are necessary.[3] In this work we identify articles that discuss the long-term complications of COVID-19.

Early in the COVID-19 pandemic, some COVID-19 patients began reporting symptoms persisting significantly past the acute phase. Finding existing supports lacking, these patients—many of whom were themselves healthcare professionals or researchers—turned online for support, naming the condition Long Covid, as a contraction of long-term COVID illness.[4] In May 2020, a patient-led group published the first survey of long-term symptoms of COVID-19.[5] Extensive subsequent research continues to show that a significant percentage of COVID-19 survivors experience ongoing multisystemic symptoms.[6–8] These symptoms include respiratory issues,

cardiovascular disease, cognitive impairment, and profound fatigue.[9–12] For many patients, these symptoms affect daily living or returning to work.[13,14] Long Covid occurs in patients with low risk of fatal outcome and in younger patients, including children.[13,15,16] Most of the morbidity burden of COVID-19 (i.e., healthy years of life lost) is in COVID-19 survivors, not fatalities.[17] Moreover, many viruses besides SARS-CoV-2—including poliovirus, varicella-zoster, Epstein-Barr, Zika, West Nile, and SARS-CoV—have been implicated in long-term sequelae.[18–23]

Long Covid remains incompletely understood, however, despite increasing evidence for several theories and notable overlaps with other conditions, including myalgic encephalomyelitis/chronic fatigue syndrome.[24,25] Reported incidence rates vary widely, from 9% to 81% according to one meta-analysis.[26] Evidence for widely effective treatments is lacking.[27] While consensus-based case definitions are emerging, definitions of Long Covid used in the literature vary substantially, which impairs building on previous work.[28–30] Nevertheless, there is increasing recognition that COVID-19 is not only a mass death event, but—through Long Covid—also a mass disabling event, making it a pressing public health concern.

Our initial analysis of the published literature found a wide variety of terms used to refer to Long Covid, but it also found that the condition is more commonly described rather than named. Querying for Long Covid articles is therefore challenging: precise queries such as "post-acute sequelae of SARS-CoV-2 infection" return limited results, whereas broad queries such as "post COVID symptoms" return many false positives.

In this work, our goal is to identify biomedical research articles relevant to Long Covid that are useful to researchers, clinicians, and patients/advocates. Theoretically, the task is a binary text classification task.[31,32] However, the objective is to comprehensively identify uncommon articles describing a novel disease entity—Long Covid—that is incompletely understood, inconsistently defined, lacks established terminology, and is frequently not named. The class imbalance and large number of articles to be classified suggest actively choosing which articles to annotate manually. We therefore employ a human-in-the-loop approach utilizing active learning to identify the articles where manual annotation would be most useful.[33] In preliminary experiments, however, we found conventional methods—uncertainty sampling with either classical machine learning or transformer-based deep learning—failed to differentiate relevant articles with language differences from the large number of irrelevant articles.

Our work therefore emphasizes thorough data exploration. We utilize multiple relevance signals as differing views of the data to identify areas of disagreement between individual signals. We also break the manually annotated data reserved for training into multiple subsets, training models on each to identify articles whose predictions are not based on robust patterns. We further utilize sources of labels available without training data, which are sometimes noisy. We combine these approaches using the weakly supervised method data programming, which integrates a set of task-specific noisy signals, called labeling functions, without additional training data.[34]

The contributions of this article are 3-fold. First, we report the creation of the Long Covid Collection, a literature resource of 8,950 articles (through July 29, 2022) relevant to an urgent public health concern. The Long Covid Collection is publicly available within the LitCovid portal, a widely used literature hub with over 290,000 articles specific to COVID-19. Second, we present an analysis of the Long Covid Collection, demonstrating that 69.0% of relevant articles do not mention Long Covid directly, making identification via query difficult. Third, we present a framework for comprehensively identifying articles relevant to concepts without established terminology, combining human-in-the-loop machine learning and data programming. We further present three extensions to data programming. We evaluate the automated prediction model on a held-out set of manually annotated articles, demonstrating a receiver operating characteristic (ROC) area under the curve (AUC) of 0.8454. We also compare our approach to several other approaches to identifying Long Covid articles, demonstrating an over 3-fold improvement in sensitivity.

RESULTS

Definition and guidelines

Following the broadest early definition with substantial support, we define Long Covid to be ongoing symptoms at least 4 weeks after initial symptoms.[35] We therefore label an article as relevant to Long Covid if it meets the following two criteria: first, the article must consider adverse effects resulting from COVID-19, i.e., SARS-CoV-2 infection. Second, the article must report outcomes or symptoms over a time frame that includes at least 4 weeks post infection. While the goal is to label articles as relevant if they contain useful discussion of the long-term symptoms caused by COVID-19, several aspects make these relevance judgements difficult. First, articles do not need to mention Long Covid by name to be relevant. Second, the symptoms may be of any type, provided they are persistent and caused by COVID-19. Third, the relevant discussion of persistent symptoms may occur in the full text rather than the abstract. Finally, articles that only refer to Long Covid in passing—such as to mention that long-term sequelae should be studied—are not relevant, even if they would be returned by keyword search. We provide the full annotation guidelines in Table S1.

We performed a small manual inter-annotator agreement study to verify the repeatability of the annotator guidelines. We randomly selected 100 of the articles previously annotated by the primary annotator (R.L., a bioinformatician with previous annotation experience). These were labeled by the senior annotator (J.W., an M.D./Ph.D.). Each article was labeled as relevant or not relevant, using the full text of the article as needed. The annotators agreed on 87 articles (61 Relevant, 26 Irrelevant) for a raw inter-annotator agreement of 87.0%. Cohen's kappa, which controls for chance agreement and ranges from −1.0 to 1.0, was 0.70, corresponding to substantial agreement.[36]

Nearly all annotator disagreements were due to the difficulty of clearly establishing the timing of the symptoms described, which often requires careful analysis. For example, the timeline for the study described in the full text for PMID: 32548209 is clearly not long enough to meet the 4 weeks required by our guidelines. While this would suggest that the article is not relevant, the article does not specify the length of time from initial infection to enrollment in the study and refers to symptoms persisting after clinical recovery. Other annotator disagreements were also primarily

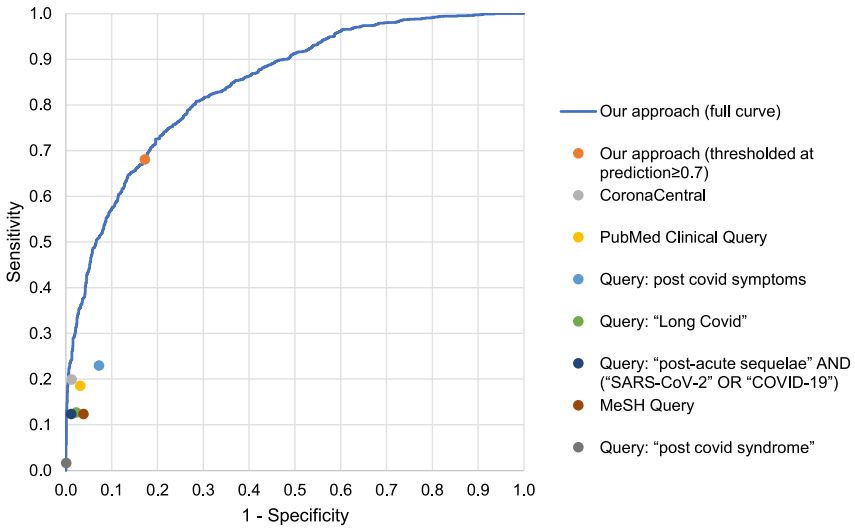


Figure 1. Receiver operating characteristic (ROC) curve of our results

Shown with the sensitivity/specificity points for our results thresholded at prediction $\geq$ 0.7 and several alternative methods of collecting articles relevant to Long Covid. The area under the curve (AUC) is 0.8454.

due to textual ambiguities; for example, PMID: 33756229 refers to patients that test positive again without clearly specifying whether the cases were due to either reactivation, which would be relevant, or reinfection, which would not be relevant.

Data summary

Our input dataset consists of LitCovid, a literature resource of COVID-19 articles in PubMed, updated daily. LitCovid categorizes each article with eight broad topics (general information, mechanism, transmission, diagnosis, treatment, prevention, case report, and epidemic forecasting). LitCovid considers all PubMed articles other than preprints. We therefore use LitCovid as a comprehensive collection of articles about COVID-19, and all articles in LitCovid were considered. The oldest articles in LitCovid were published in January 2020.

We created the Long Covid Collection using a human-in-the-loop machine learning process with the goal of minimizing human effort while creating a classifier that is both accurate and able to identify articles requiring human labels due to uncertainty. In our usage, an active learning process iteratively chooses articles for the human annotators to judge for relevance, which are then used to improve an automated system. The updated system is then used to select a new set of articles for annotation, focusing on articles where the automated system is uncertain, and the process repeats. The automated system is designed as an ensemble of lightweight, independent, classifiers with differing views of the data. This design is critical for focusing the human annotation effort on articles where the model is uncertain. Our iterative process provides two features not available with a more conventional approach: first, it produces a high probability of identifying all articles relevant to Long Covid, and second, it uses human annotation effort more efficiently.

The Long Covid Collection was first released on August 1, 2021, consisting of 2,056 articles, and it is updated weekly. As of July 29, 2022, the Long Covid Collection contained 8,950 articles, gaining approximately 133 articles per week on average. Approximately 2.9% of articles in LitCovid are relevant, substantially skewing the relevant and irrelevant classes. Our annotation process prioritizes articles where the automated system is uncertain, and this high skew results in the least certain articles containing a high proportion of relevant articles. As a result, the manually annotated articles contain a greater number of relevant articles than irrelevant articles. Moreover, the irrelevant articles manually annotated tend to be those that are difficult to distinguish from relevant articles. As of July 29, 2022, there were 10,149 manually annotated articles, 5,800 annotated as relevant, and 4,349 annotated as irrelevant. As new articles are annotated manually, one-quarter of them are randomly reserved for validation.

Validation: Comparison methods and evaluation

We compared our results to several other collections on Long Covid. The CoronaCentral resource contains articles related to several coronaviruses, including SARS-CoV-2, with automated predictions for both topics and various entities.[37] We consider articles annotated with both SARS-CoV-2 and the Long Haul topic as Long Covid articles according to CoronaCentral, using CoronaCentral version 84. PubMed Clinical Queries uses predefined keyword filters to help users perform and refine specialized searches. The queries for COVID-19 are intended to limit results to articles on SARS-CoV-2 with a particular topic; we use the Long COVID filter, which is implemented as a keyword query and listed in full in Note S1. Medical Subject Headings (MeSH) is a controlled vocabulary used for indexing articles by topic.[38] We created a Long Covid query from MeSH terms by combining COVID-19 or SARS-CoV-2 with terms reflecting the post-acute phase, also listed in full in Note S1. Finally, we created several textual queries from the most common Long Covid terms. These queries are as follows: "post covid symptoms," "Long Covid," "post-acute sequelae" AND ("SARS-CoV-2" OR "COVID-19"), and "post covid syndrome."

The evaluation set, created by randomly reserving one-quarter of all articles annotated manually, contains 1,450 positive articles and 1,088 negative articles. We evaluate the results using sensitivity and specificity, which can be visualized using the ROC curve and can be summarized as the AUC.[39] Since the comparison approaches provide only binary predictions, we binarize our results for comparison by thresholding at a prediction of 0.7. The ROC curve for our results and the sensitivity/specificity points for all comparison approaches can be seen in Figure 1. The AUC is 0.8454. Ablating any individual labeling function types (with one exception) results in a small or negligible change to the AUC, demonstrating that our framework allows the overall model to be robust for removing individual sources (see Note S2). The one exception is LitSuggest (trained on manual annotations), which reduces the AUC by 0.1030, to 0.7424.

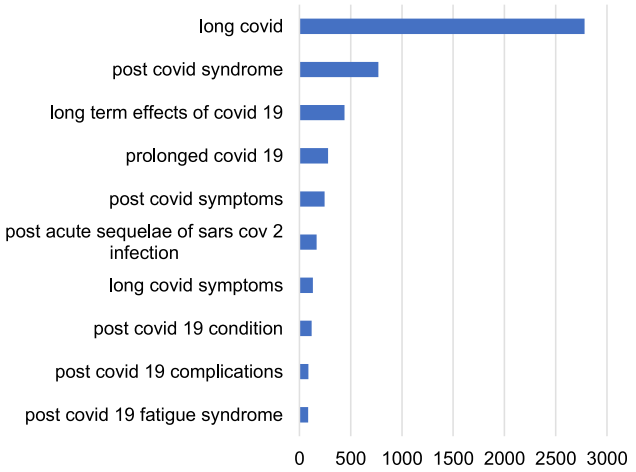


Figure 2. Terms for Long Covid found most frequently by the Long Covid grammar

The grammar found a total of 7,378 mentions of Long Covid, representing 763 unique phrases, ignoring capitalization and punctuation.

While the comparison methods provide high specificity, the highest sensitivities are for the post-covid symptoms query (0.2297) and CoronaCentral (0.1993), whereas the sensitivity of our thresholded results is 0.6807.

Resource analysis

We created a grammar-based named entity recognizer to identify mentions of Long Covid. This grammar is designed to accommodate significant language variability and extends a grammar previously created to identify mentions of COVID-19 and SARS-CoV-2.[40] The Long Covid grammar-based named entity recognizer identified 7,378 mentions of Long Covid, representing 763 unique phrases (after normalizing case and punctuation); the most frequent are summarized in Figure 2.

Despite the flexibility of the grammar, 69.0% of the articles in the Long Covid Collection do not contain an identifiable term for Long Covid. This is commonly caused by the article referring to Long Covid using a description rather than a term. While grammar can identify many descriptive phrases, such as "long-term outcomes of COVID-19," some of the descriptions used by authors to refer to Long Covid remain beyond the ability of the grammar to recognize. This is often due to some qualification, such as an anatomical system. For example, the phrase "residual respiratory impairment after COVID-19 pneumonia" (PMID: 34273962) strongly suggests a respiratory form of Long Covid but could not be recognized by the grammar. A more advanced recognition technique should be able to recover some additional descriptive mentions, but any such mentions remaining are not common. A more advanced technique may also help reduce false positives in the grammar, though these are quite rare. For example, "… how long COVID-19 (SARS-CoV-2) survives …" (PMID: 32967479) includes the phrase "long COVID" but does not refer to Long Covid.

Inspection of the term frequencies, as seen in Figure 2 and in the full data, shows that the frequency of Long Covid terms reflect a long tail distribution. Plotting the rank of each term against its frequency in a log-log plot results in an approximately straight line (data not shown), indicating a Zipf distribution, as is common for linguistic data.[41] Identifying Long Covid articles by identifying synonymous terms is therefore subject to diminishing returns, and additional methods are required.

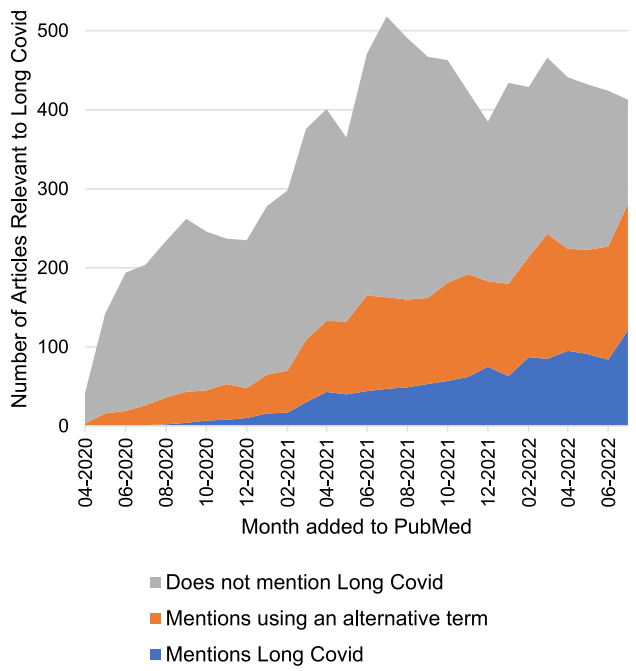


Figure 3. Terms used to refer to Long Covid over time

Articles that mention Long Covid use the name Long Covid at least once. Articles that use an alternative term mention Long Covid at least once via identifiable synonym. Articles that do not mention Long Covid do not contain an identifiable term for Long Covid. All articles listed are relevant to Long Covid.

Naming trend for Long Covid over time

Figure 3 shows the naming trend over time, with all articles relevant to Long Covid listed as either mentioning Long Covid directly (i.e., using the term Long Covid), mentioning Long Covid but using a different term, or not mentioning Long Covid by name. All articles listed are relevant to Long Covid, and each article is only counted once.

We see that not only is Long Covid the most common term used in the literature to refer to Long Covid (see Figure 2), but its use also appears to be increasing slowly. Moreover, the percentage of articles that are relevant to Long Covid but do not refer to it using an identifiable term appears to be decreasing. However, these changes appear to be gradual, suggesting that the lack of terminological consensus will remain for some time. Unfortunately, this reluctance to name the condition likely makes it more difficult for consensus to build: articles that rely on descriptions will be more difficult to locate since automated recognition of descriptions is known to be much more difficult than names.[42]

Analysis of entities mentioned

PubTator is a web-based system providing annotations for six entity types: genes/proteins, genetic variants, diseases, chemicals, species, and cell lines.[43] We compared the annotation

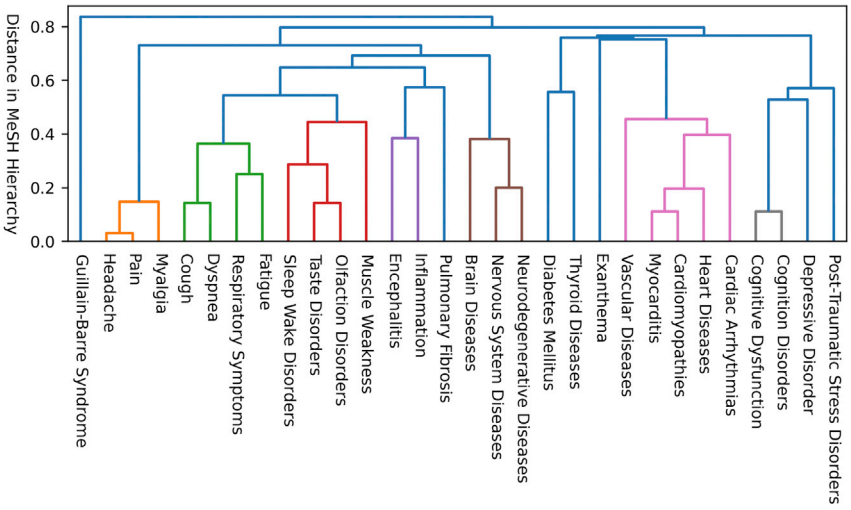


Figure 4. Dendrogram of the disorders most frequently mentioned in the Long Covid Collection

Disorders are filtered if their annotation rate is less than in the general COVID-19 literature ($p < 0.01$, Fisher exact test). Disorders are clustered according to the number of ancestors in common in the Medical Subject Headings (MeSH) hierarchy.

rate for entities annotated by PubTator in the Long Covid Collection and LitCovid. We found that the entities that showed a statistically significant difference ($p < 0.01$, Fisher exact test) were primarily disorders that appear more frequently in the Long Covid Collection than in the general COVID-19 literature. We selected for these and further removed disorders specific to COVID-19, such as multisystem inflammatory syndrome in children (MIS-C). Figure 4 visualizes the 30 disorders that appear most frequently in the Long Covid Collection in a dendrogram, clustered according to the number of ancestors in common in the MeSH hierarchy.

Figure 4 underscores the great variety of body systems affected by Long Covid, demonstrating that Long Covid is a multisystemic condition. Several of the symptoms most common in Long Covid patients are listed, such as fatigue and cognitive dysfunction.[6] Neurological and cardiovascular conditions also appear prominently. A notable trend that is less apparent in this clustering are conditions due to immune system dysregulation, such as Guillain-Barre syndrome, myocarditis (inflammation of the heart muscle), encephalitis (inflammation of the brain), and inflammation itself. Interestingly, several symptoms closely associated with COVID-19 are seen to appear more frequently in connection with Long Covid than COVID-19, such as respiratory symptoms, dyspnea, and olfaction disorders.

The chemicals that appear statistically significantly more frequently in the Long Covid Collection than in LitCovid include several classes of drugs. These are steroids (prednisolone, methylprednisolone), NSAIDs (aspirin, indomethacin), and anti-fungals (amphotericin B). Other chemicals that appear more frequently in the Long Covid Collection include chemicals used in various clinical tests (gadolinium, fluorodeoxyglucose F18, carbon monoxide) and cortisone, due to its relation to adrenal insufficiency.

Analysis of topic clusters

We use the probabilistic distributional clustering (PDC) algorithm to identify topics within the Long Covid Collection.[44] PDC uses terms, phrases, and MeSH terms occurring within a collection as input and utilizes their probability of co-occurrence to partition the set of input features into disjoint groups. Documents can then be scored with respect to each topic identified and may receive a high score for more than one topic. Figure 5 shows the most frequent topics identified by the PDC clustering algorithm, organized into four aspects: study methods, interventions (treatments and tests), systemic dysfunctions, and specific disorders. Articles that are identified as containing multiple topics within a chart contribute fractionally to each topic, so that each article is counted only once. The names for each topic are manually generated but represent the most common phrases in the topic.

Broadly, the topics show an expansion through mid-2021, with a slight contraction after. Figure 5A shows significant expansions for both cohort studies and systematic reviews, while case studies show a slight contraction since late 2021. This pattern suggests increasing scientific rigor over time.[45] Interventions, in Figure 5B, include both testing and treatments but were not common. Moreover, the interventions that are seen are not specific to Long Covid, but rather broad care classifications or extensions of COVID-19 interventions. Figure 5C describes systemic dysfunctions; we see that the initial discourse was dominated by pulmonary dysfunction, though discussion of neurological and cognitive dysfunction topic also began early. All other systemic dysfunction topics start small but increase over time, indicating increasing recognition of the long-term effects of COVID-19 on multiple body systems. Specific disorders appear in Figure 5D. These disorders affect a wide variety of body systems, and—with one clear exception—all follow the general trend of increasing counts. The exception, the viral pneumonia topic, is the only specific disorder topic that starts with considerable counts, then contracts significantly at the end of 2020. This trend shows that acute COVID-19, which primarily causes viral pneumonia, is typically no longer discussed with Long Covid: Long Covid is now discussed as a separate entity.

DISCUSSION

COVID-19 has caused widespread mortality and has strained healthcare systems worldwide.[46] Estimates of the overall morbidity burden show, however, that most of the burden lies in COVID-19 survivors.[17] While estimates of the prevalence of Long Covid vary, a recent meta-analysis shows that the prevalence of symptoms beyond 4 weeks is quite high: 43%.[26] Moreover, the effects experienced many years or decades later are yet unknown. Continuing research into Long Covid is critical, and identifying Long Covid articles comprehensively allows the articles to be analyzed as a set. Our analyses of both the entity mentions and the topic clustering support the view that Long

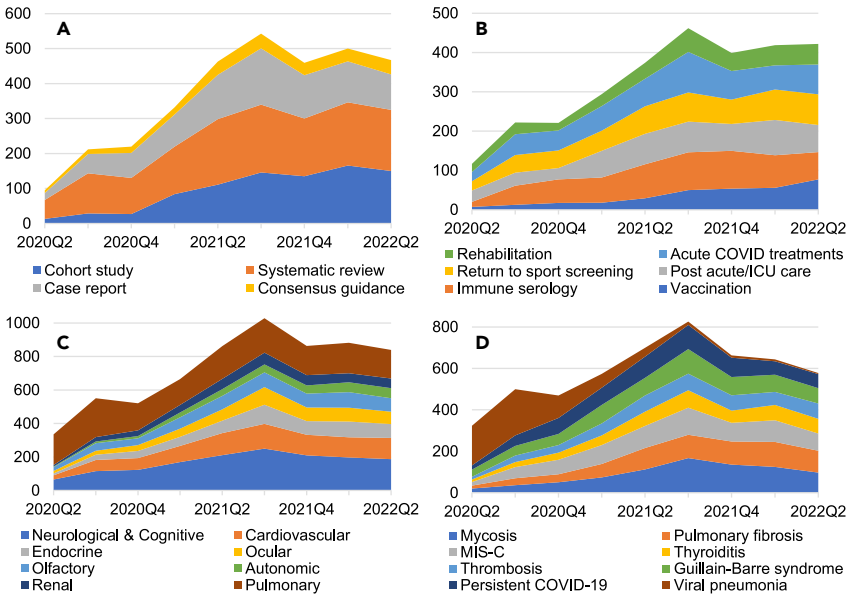


Figure 5. Most frequent topics in the Long Covid Collection over time

(A–D) Topic names are manually generated but reflect the most common phrases in the topic. Study types (A) show increased rigor over time, while interventions (B) show a lack of treatments or tests specific to Long Covid. Long Covid is a complex, multisystemic, condition that causes a wide variety of potentially serious systemic dysfunctions (C) and specific disorders (D).

Covid is a multisystemic condition. Notably, our experiments did not uncover support for psychological causes. However, both the analysis of entity mentions and the topic clustering find a greater number of symptoms associated with Long Covid than treatments for those symptoms. Moreover, analysis of the LitCovid topics (see Figure S1) shows that articles relevant for Long Covid are far more likely to be case reports than other articles on COVID-19, and the number of articles discussing mechanism has remained relatively stable over time, even as the diagnosis and treatment topics have expanded.

We also find that there is a significant overlap between the symptoms associated with COVID-19 and the symptoms associated with Long Covid. This supports the view that the name Long Covid is descriptive of the condition as extended COVID. The fact that the name Long Covid is both the most common name and an increasing number of articles use the name suggests that a consensus is building, albeit slowly. However, this does not preclude another name—or subtype name—gaining favor once the etiology—or etiologies—are identified. Interestingly, other names were initially used for both COVID-19 and SARS-CoV-2, pneumonia of unknown etiology and 2019-nCoV, respectively, though a consensus built fairly quickly.[40] Unlike COVID-19 and SARS-CoV-2, however, standardizing organizations have not argued for specific terminology for Long Covid; use of the term Long Covid is largely due to patient advocacy efforts.[4]

Our definition of Long Covid is primarily time-based, while many of the studies in the literature that discuss sequelae of COVID-19 are primarily concerned with a specific body system. This is particularly the case with neurological effects, which are prevalent in Long Covid but may also appear much earlier, even as the initial manifestation of infection. Our annotators noted that many articles could not be labeled from just the title and abstract; this is primarily because of the condition that symptoms must be present at least a month after initial infection. Moreover, our combination of a broad definition and short time frame (1 month post infection) implies the inclusion of some post-COVID conditions with an acute presentation, notably MIS-C and mucormycosis. However, these conditions do share some elements, such as immune system dysregulation. Nevertheless, Long Covid remains an area of active research and updates are expected.

Our work has several limitations. Our analyses are mostly correlational and do not show, for example, that the association between specific symptoms are definitively caused by Long Covid. Our method does not specifically address language drift, though the effects of language drift should be ameliorated somewhat through the iterative annotation process. We anticipate that the Long Covid Collection itself will remain relevant for some time even if a strong consensus builds and the level of terminological variation drops significantly. Again, Long Covid remains an area of active research, and new developments are expected.

Our methods are automated and do not produce perfect accuracy. However, this is partially due to inherent ambiguities. For example, it is sometimes difficult to label articles based on the title and abstract, such as determining whether a reference to COVID-19 patients refers to patients who had COVID-19 previously or patients who currently have acute COVID-19 (e.g., PMID: 35043098). Unfortunately, the full text is often not available.

It is difficult to provide a definitive discussion of how many articles must be annotated to provide high coverage. However, our framework already provides an AUC of approximately 0.70 using only the labeling functions that do not require training (data not shown). We performed a series of experiments iteratively rerunning article selection and found that 1,000 manually annotated articles reliably produces a model with an AUC of over 0.80 (data not shown). In this work we also intended to provide a reliable, comprehensive collection of articles on Long Covid; we therefore proceeded to manually annotate most of the relevant articles and a nearly equal number of irrelevant articles. Note, however, that this still results in a significant annotation savings compared with manual annotation: we annotated approximately 3.7% of the articles in LitCovid, representing an annotations savings of 96.3%. We therefore believe that our framework reduces the need for manual annotations when identifying high-variation terminology.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to the lead contact, Zhiyong Lu (zhiyong.lu@nih.gov).

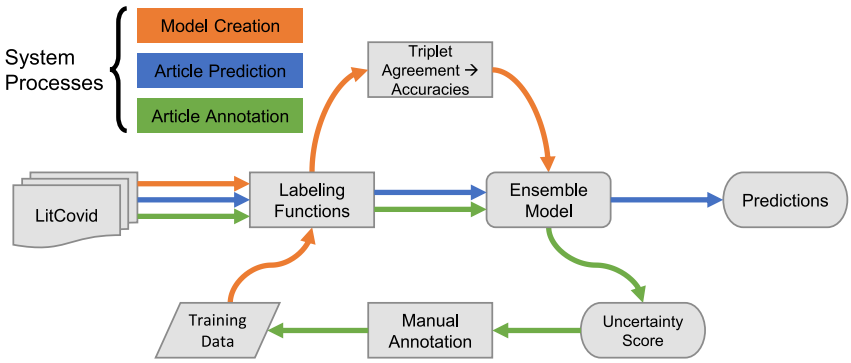


Figure 6. System overview

System diagram illustrates the flow of data for the three primary system processes: model creation, article prediction, and article annotation.

Materials availability

This study did not generate any new unique materials.

Data and code availability

- Original data (article classifications, manual annotations, and mentions) have been deposited at Zenodo under https://doi.org/10.5281/zenodo.7308463 and are publicly available as of the date of publication.
- This paper analyzes existing, publicly available data. The Zenodo DOI for CoronaCentral is https://doi.org/10.5281/zenodo.6896953. The API URL for LitCovid is https://www.ncbi.nlm.nih.gov/research/coronavirus-api/export/all/tsv. The API URL for PubTator is https://www.ncbi.nlm.nih.gov/research/pubtator/api.html.
- All original code has been deposited at Zenodo under https://doi.org/10.5281/zenodo.7308820 and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Human-in-the-loop process overview

Our goal is to comprehensively identify all PubMed articles relevant to Long Covid. These articles are a subset of articles relevant to COVID-19, which are already collected by the LitCovid resource. The task therefore becomes classifying each article in LitCovid as relevant or irrelevant to Long Covid. However, since the objective is to identify articles relevant to a novel disease entity that is incompletely understood, inconsistently defined, lacks established terminology, and is frequently not named, our approach prioritizes data exploration in addition to prediction.

In human-in-the-loop machine learning, data points can be selected for human annotation based on either diversity sampling or uncertainty sampling.[33] In diversity sampling, the data are clustered and instances are chosen to ensure that each cluster is represented. Uncertainty sampling, on the other hand, prioritizes instances closest to the decision boundary or instances with the largest variation in predictions. Our initial analysis showed fast improvements with uncertainty sampling, with no additional benefit with diversity sampling. Our framework therefore prioritizes articles for annotation when close to the decision boundary or when the automated predictions show high variation. In preliminary experiments, we found conventional methods—uncertainty sampling with either classical machine learning or transformer-based deep learning—failed to differentiate relevant articles with language differences from the large number of irrelevant articles.

The automated system therefore employs a semi-supervised approach, allowing predictions from multiple relevance signals. The system combines these signals—some of which are created with supervised classification—using the weakly supervised framework data programming.[34] These relevance signals, called labeling functions, are derived from disparate data sources, producing multiple views of the data that are sometimes contradictory. Data programming uses the labeling functions to create an ensemble model without training data.[34] Triplet methods are a recent development in data programming that provide a closed form solution that does not require iterative training.[47] We extend triplet data programming to allow probabilistic labeling functions (rather than only binary), to improve reliability of the ensemble model and to provide uncertainty scores.

We provide an overview of our framework in Figure 6, which illustrates the data flow for the three high level processes used by our system. The first process, model creation, prepares the labeling functions (some of which require training data), retrieves a label for each article from each labeling function, and creates the ensemble model using data programming. The second process, article prediction, uses the model to predict the relevance of every article. The third process, article annotation, uses the model to identify uncertain predictions; articles with high uncertainty are then prioritized for manual annotation. One-quarter of annotated articles are reserved for evaluation, and the remainder are added to the training data. The PubMed query used to initialize the iterative annotation process was "long covid" OR ("sequelae" AND ("COVID-19" OR "SARS-CoV-2")), and article annotation is performed using the LitSuggest online interface, https://www.ncbi.nlm.nih.gov/research/litsuggest/.

Data programming

While supervised machine learning requires abundant training data, data programming creates a model by aggregating weaker forms of supervision.[34] This creates an ensemble model, similar to the machine learning method stacked generalization, also known as stacking.[48] In the data programming paradigm, the practitioner creates labeling functions—task-specific functions that imperfectly label instances—instead of labeling instances. Labeling functions may take many forms, including rule-based patterns, dictionary lookups, and noisy supervised classifiers. The labeling functions are applied to a large amount of unlabeled data, and the agreement rates between the labeling functions are then used to infer the accuracy of each labeling function. The accuracies for each labeling function then form the parameters of a generative model that can be used to label new data points, applying a small amount of knowledge—in the form of labeling functions—to accurately label a large amount of data. Note that various forms of noise—including missing values and disagreements—are anticipated and handled by the framework.

Since human-in-the-loop methods begin with very little labeled data, applying data programming within a human-in-the-loop approach would seem to be ideal. However, the human-in-the-loop approach requires the automated system to be repeatedly retrained, which is inconvenient despite the simplicity of the generative model due to the large data sizes involved. Data programming with triplet methods makes repeated retraining unnecessary by directly estimating the accuracy of each labeling function using a closed form solution.[47] This solution requires only calculating the pairwise agreement rates between a triplet of labeling functions; it can be extended to an arbitrary number of labeling functions by iterating through all possible triplets and averaging the results.

Specifically, given a triplet of binary labeling functions $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ of the form $\mathcal{L}(x) \in \{0, 1\}$, which are conditionally independent given the class, the estimated accuracy of $\mathcal{L}_1$ is

$$estimated\_accuracy(\mathcal{L}_1) = \frac{1}{2}\sqrt{\frac{agree(\mathcal{L}_1,\mathcal{L}_2)\cdot agree(\mathcal{L}_1,\mathcal{L}_3)}{agree(\mathcal{L}_2,\mathcal{L}_3)}} + \frac{1}{2}$$

Where, we define *agree* for a pair of labeling functions as

$$agree(\mathcal{L}_1,\mathcal{L}_2) = \frac{1}{|X|}\sum_{x \in X} equal(\mathcal{L}_1(x),\mathcal{L}_2(x))$$

And, we define *equal* as

$$equal(\mathcal{L}_1(x),\mathcal{L}_2(x)) = \begin{cases} 1, \text{ if } \mathcal{L}_1(x) = \mathcal{L}_2(x) \\ -1, \text{ if } \mathcal{L}_1(x) \neq \mathcal{L}_2(x) \end{cases}$$

Table 1. Labeling functions used to identify articles relevant to Long Covid

Name	Description	Requires training data
Long Covid grammar	A purpose-built grammar-based named entity recognition system to identify mentions of Long Covid. Uses the full text, if available. See Table S2	no
LitSuggest, trained by query	Predictions from the LitSuggest web-based literature curation tool,[50] trained using a query for positives and random articles for negatives	no
LitSuggest, trained on annotations	Predictions from the LitSuggest web-based literature curation tool,[50] trained using the annotated training data	yes
PubTator entity annotations	Entities from the PubTator annotation system,[43] using the full text, if available	yes
MeSH headings	Medical subject headings indexed by the National Library of Medicine indexing team, if available	yes
CoronaCentral Long-Haul topic	Articles from the CoronaCentral portal,[37] annotated with the Long Haul topic and one or more mentions of SARS-CoV-2	no
CoronaCentral entity annotations	Entity annotations provided by the CoronaCentral portal	yes
Bias	Labels all articles as probably not relevant	no

Complete descriptions of each labeling function are provided in the supplemental experimental procedures.

Note that in the previous definition the labeling functions $\mathcal{L}(x)$ have binary values. We extend the accuracy calculation to handle probabilistic labeling functions $\widehat{\mathcal{L}}(x) \in [0, 1]$. We first define $sample(x)$ to be a function that discretizes its input $x \in [0, 1]$ by returning the value 1 with probability x and returning 0 with probability $1 - x$. We can then define the *equal* function for probabilistic labeling functions by averaging over many samples, each of which has discrete binary values. We note that the result converges to a closed form solution as the number of samples approaches infinite, specifically the following:

$$equal(\widehat{\mathcal{L}}_1(x), \widehat{\mathcal{L}}_2(x)) = \lim_{n\to\infty} \frac{1}{n} \sum_n equal(sample(\widehat{\mathcal{L}}_1(x)), sample(\widehat{\mathcal{L}}_2(x)))$$

$$= (2\widehat{\mathcal{L}}_1(x) - 1) \cdot (2\widehat{\mathcal{L}}_2(x) - 1)$$

Also note that under this formulation, a labeling function that cannot provide a prediction for any given input may abstain by returning exactly 0.5.

The accuracy of each labeling function can be estimated by forming a triplet with any other two labeling functions. Since there is an abundance of labeling functions, we gather many estimates for each labeling function by using all available pairs and use their mean as the final accuracy estimate. However, we improve the reliability of the accuracy estimates by ignoring pairs of labeling functions whose agreement may be due to chance. We model the agreement between a pair of labeling functions as a binomial distribution, where the number of trials is the number of instances, and the number of successes is the number of instances where the labeling functions agree. We calculate a confidence interval on the agreement rate between each pair of labeling functions and discard the pairs where the confidence interval includes 0.5. Our implementation uses the Wilson method at the 95% confidence level.[49]

Prediction and uncertainty sampling

Given the vector of accuracies for each labeling function, a, and the vector l for a given article, the prediction $p \in [0, 1]$ is

$$p = \frac{1}{1+e^{-x}}, \text{ where } x = \sum_i (2l_i - 1) \times \log \frac{a_i}{1 - a_i}$$

We recalibrate the predictions of the generative model using the mean and standard deviations of the articles manually annotated positively and negatively.

In a human-in-the-loop approach, the automated model provides both predictions as well as prioritizing instances for manual annotation.[33] We prioritize articles for manual annotation by identifying articles whose predictions are uncertain. We use two approaches for uncertainty sampling, specifically distance to threshold and prediction variation. Under distance to threshold, instances with predictions closer to the decision boundary have higher uncertainty; given a prediction p and a threshold t, the distance to the threshold is $dist = abs(p - t)$. We support uncertainty sampling via variation by running the data programming inference multiple times, masking 50% of the labeling functions during each run, then calculating the inter-quartile range of the predictions for each instance (iqr). Our final selection criterion combines the distance to the threshold and variation calculations, choosing the unlabeled instances that simultaneously minimize $dist$ and maximize iqr.

Description of labeling functions

Data programming with triplet methods requires the assumption that the labeling functions are conditionally independent given the class.[47] We use eight types of labeling functions, chosen for providing complementary views. Table 1 describes the labeling functions briefly; they are fully described in the supplemental experimental procedures. Several of the labeling function types (LitSuggest, MeSH headings, entity annotations) require training data. Manually annotated data are split into four parts: one part is reserved for evaluation. The remainder are used to train three independent labeling functions for each labeling function type.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100659.

ACKNOWLEDGMENTS

This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

AUTHOR CONTRIBUTIONS

Conceptualization: R.L. and Z.L.; overall methodology, software, and analysis: R.L.; supervision: Z.L.; annotation guidelines: R.L., R.I., J.W., A.A., and Q.C.; data curation and validation: R.L. and J.W.; PDC analysis: R.I.; LitCovid resources and software: A.A. and Q.C.; LitSuggest software: A.A.; drafted initial manuscript: R.L.; all authors reviewed, edited, and approved of the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 29, 2022
Revised: September 19, 2022
Accepted: November 17, 2022
Published: December 1, 2022

REFERENCES

1. Chen, Q., Allot, A., and Lu, Z. (2021). LitCovid: an open database of COVID-19 literature. Nucleic Acids Res. *49*, D1534–D1540. https://doi.org/10.1093/nar/gkaa952.
2. Chen, Q., Allot, A., and Lu, Z. (2020). Keep up with the latest coronavirus research. Nature *579*, 193. https://doi.org/10.1038/d41586-020-00694-1.
3. Chen, Q., Allot, A., Leaman, R., Wei, C.-H., Aghaarabi, E., Guerrerio, J., Xu, L., and Lu, Z. (2022). LitCovid in 2022: an information resource for the COVID-19 literature. Nucleic Acids Res. *2022*, gkac1005. https://doi.org/10.1093/nar/gkac1005.
4. Callard, F., and Perego, E. (2021). How and why patients made Long Covid. Soc. Sci. Med. *268*, 113426. https://doi.org/10.1016/j.socscimed.2020.113426.
5. Patient Led Research Collaborative (2020). Report: what does COVID-19 recovery actually look like? An analysis of the prolonged COVID-19 symptoms survey by patient-led research team. https://patientresearchcovid19.com/research/report-1/.
6. Davis, H.E., Assaf, G.S., McCorkell, L., Wei, H., Low, R.J., Re'em, Y., Redfield, S., Austin, J.P., and Akrami, A. (2021). Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. EClinicalMedicine *38*. https://doi.org/10.1016/j.eclinm.2021.101019.
7. Nalbandian, A., Sehgal, K., Gupta, A., Madhavan, M.V., McGroder, C., Stevens, J.S., Cook, J.R., Nordvig, A.S., Shalev, D., Sehrawat, T.S., et al. (2021). Post-acute COVID-19 syndrome. Nat. Med. *27*, 601–615. https://doi.org/10.1038/s41591-021-01283-z.
8. Taquet, M., Dercon, Q., Luciano, S., Geddes, J.R., Husain, M., and Harrison, P.J. (2021). Incidence, co-occurrence, and evolution of long-COVID features: a 6-month retrospective cohort study of 273,618 survivors of COVID-19. PLoS Med. *18*, e1003773. https://doi.org/10.1371/journal.pmed.1003773.
9. Hayes, L.D., Ingram, J., and Sculthorpe, N.F. (2021). More than 100 persistent symptoms of SARS-CoV-2 (long COVID): a scoping review. Front. Med. *8*, 750378. https://doi.org/10.3389/fmed.2021.750378.
10. Xie, Y., Xu, E., Bowe, B., and Al-Aly, Z. (2022). Long-term cardiovascular outcomes of COVID-19. Nat. Med. *28*, 583–590. https://doi.org/10.1038/s41591-022-01689-3.
11. Douaud, G., Lee, S., Alfaro-Almagro, F., Arthofer, C., Wang, C., McCarthy, P., Lange, F., Andersson, J.L.R., Griffanti, L., Duff, E., et al. (2022). SARS-CoV-2 is associated with changes in brain structure in UK Biobank. Nature *604*, 697–707. https://doi.org/10.1038/s41586-022-04569-5.
12. Carfì, A., Bernabei, R., and Landi, F.; Gemelli Against COVID-19 Post-Acute Care Study Group (2020). persistent symptoms in patients after acute COVID-19. JAMA *324*, 603–605. https://doi.org/10.1001/jama.2020.12603.
13. Havervall, S., Rosell, A., Phillipson, M., Mangsbo, S.M., Nilsson, P., Hober, S., and Thålin, C. (2021). Symptoms and functional impairment assessed 8 Months after mild COVID-19 among health care workers. JAMA *325*, 2015–2016. https://doi.org/10.1001/jama.2021.5612.
14. Sivan, M., Parkin, A., Makower, S., and Greenwood, D.C. (2021). Post-COVID Syndrome symptoms, functional disability and clinical severity phenotypes in hospitalised and non-hospitalised individuals: a cross-sectional evaluation from a community COVID rehabilitation service. J. Med. Virol. *94*, 1419–1427. https://doi.org/10.1002/jmv.27456.
15. Blomberg, B., Mohn, K.G.-I., Brokstad, K.A., Zhou, F., Linchausen, D.W., Hansen, B.-A., Lartey, S., Onyango, T.B., Kuwelker, K., Sævik, M., et al. (2021). Long COVID in a prospective cohort of home-isolated patients. Nat. Med. *27*, 1607–1613. https://doi.org/10.1038/s41591-021-01433-3.
16. McFarland, S., Citrenbaum, S., Sherwood, O., van der Togt, V., and Rossman, J.S. (2022). Long COVID in children. Lancet Child Adolesc. Health *6*, e1. https://doi.org/10.1016/S2352-4642(21)00338-2.
17. Smith, M.P. (2021). Estimating total morbidity burden of COVID-19: relative importance of death and disability. J. Clin. Epidemiol. *142*, 54–59. https://doi.org/10.1016/j.jclinepi.2021.10.018.
18. Aston, J.W., Jr. (1992). Post-polio syndrome. An emerging threat to polio survivors. Postgrad. Med. *92*, 249–256. 260. https://doi.org/10.1080/00325481.1992.11701402.
19. Freer, G., and Pistello, M. (2018). Varicella-zoster virus infection: natural history, clinical manifestations, immunity and current and future vaccination strategies. New Microbiol. *41*, 95–105.
20. Bjornevik, K., Cortese, M., Healy, B.C., Kuhle, J., Mina, M.J., Leng, Y., Elledge, S.J., Niebuhr, D.W., Scher, A.I., Munger, K.L., and Ascherio, A. (2022). Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. Science *375*, 296–301. https://doi.org/10.1126/science.abj8222.
21. Brasil, P., Pereira, J.P., Jr., Moreira, M.E., Ribeiro Nogueira, R.M., Damasceno, L., Wakimoto, M., Rabello, R.S., Valderramos, S.G., Halai, U.A., Salles, T.S., et al. (2016). Zika virus infection in pregnant women in Rio de Janeiro. N. Engl. J. Med. *375*, 2321–2334. https://doi.org/10.1056/NEJMoa1602412.
22. Patel, H., Sander, B., and Nelder, M.P. (2015). Long-term sequelae of West Nile virus-related illness: a systematic review. Lancet Infect. Dis. *15*, 951–959. https://doi.org/10.1016/S1473-3099(15)00134-6.
23. Moldofsky, H., and Patcai, J. (2011). Chronic widespread musculoskeletal pain, fatigue, depression and disordered sleep in chronic post-SARS syndrome; a case-controlled study. BMC Neurol. *11*, 37. https://doi.org/10.1186/1471-2377-11-37.
24. Proal, A.D., and VanElzakker, M.B. (2021). Long COVID or post-acute sequelae of COVID-19 (PASC): an overview of biological factors that may contribute to persistent symptoms. Front. Microbiol. *12*, 698169. https://doi.org/10.3389/fmicb.2021.698169.
25. Kedor, C., Freitag, H., Meyer-Arndt, L., Wittke, K., Hanitsch, L.G., Zoller, T., Steinbeis, F., Haffke, M., Rudolf, G., Heidecker, B., et al. (2022). A prospective observational study of post-COVID-19 chronic fatigue syndrome following the first pandemic wave in Germany and biomarkers associated with symptom severity. Nat. Commun. *13*, 5104. https://doi.org/10.1038/s41467-022-32507-6.
26. Chen, C., Haupert, S.R., Zimmermann, L., Shi, X., Fritsche, L.G., and Mukherjee, B. (2022). Global prevalence of post-coronavirus disease 2019 (COVID-19) condition or long COVID: a meta-analysis and systematic review. J. Infect. Dis. *226*, 1593–1607. https://doi.org/10.1093/infdis/jiac136.
27. Yong, S.J. (2021). Long COVID or post-COVID-19 syndrome: putative pathophysiology, risk factors, and treatments. Infect. Dis. (Lond) *53*, 737–754. https://doi.org/10.1080/23744235.2021.1924397.
28. Akbarialiabad, H., Taghrir, M.H., Abdollahi, A., Ghahramani, N., Kumar, M., Paydar, S., Razani, B., Mwangi, J., Asadi-Pooya, A.A., Malekmakan, L., and Bastani, B. (2021). Long COVID, a comprehensive systematic scoping review. Infection *49*, 1163–1186. https://doi.org/10.1007/s15010-021-01666-x.

29. Deer, R.R., Rock, M.A., Vasilevsky, N., Carmody, L., Rando, H., Anzalone, A.J., Basson, M.D., Bennett, T.D., Bergquist, T., Boudreau, E.A., et al. (2021). Characterizing long COVID: deep phenotype of a complex condition. EBioMedicine *74*, 103722. https://doi.org/10.1016/j.ebiom.2021.103722.

30. Soriano, J.B., Murthy, S., Marshall, J.C., Relan, P., and Diaz, J.V.; W. H. O. Clinical Case Definition Working Group on Post-COVID-19 Condition (2021). A clinical case definition of post-COVID-19 condition by a Delphi consensus. Lancet Infect. Dis. *22*, e102–e107. https://doi.org/10.1016/S1473-3099(21)00703-9.

31. Chen, Q., Allot, A., Leaman, R., Islamaj, R., Du, J., Fang, L., Wang, K., Xu, S., Zhang, Y., Bagherzadeh, P., et al. (2022). Multi-label classification for biomedical literature: an overview of the BioCreative VII LitCovid Track for COVID-19 literature topic annotations. Database *2022*, baac069. https://doi.org/10.1093/database/baac069.

32. Bangyal, W.H., Qasim, R., Rehman, N.U., Ahmad, Z., Dar, H., Rukhsar, L., Aman, Z., and Ahmad, J. (2021). Detection of fake news text classification on COVID-19 using deep learning approaches. Comput. Math. Methods Med. *2021*, 5514220. https://doi.org/10.1155/2021/5514220.

33. Monarch, R.M. (2021). Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI (Manning Publications).

34. Ratner, A., Sa, C.D., Wu, S., Selsam, D., and Ré, C. (2016). Data programming: creating large training sets, quickly. Adv. Neural Inf. Process. Syst. *29*, 3567–3575.

35. Centers for Disease Control and Prevention (2020). Post-COVID conditions: information for healthcare providers. https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/post-covid-conditions.html.

36. Landis, J.R., and Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics *33*, 159–174.

37. Lever, J., and Altman, R.B. (2021). Analyzing the vast coronavirus literature with CoronaCentral. Proc. Natl. Acad. Sci. USA *118*. e2100766118. https://doi.org/10.1073/pnas.2100766118.

38. Lipscomb, C.E. (2000). Medical subject headings (MeSH). Bull. Med. Libr. Assoc. *88*, 265–266.

39. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recogn. Lett. *27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

40. Leaman, R., and Lu, Z. (2020). A comprehensive dictionary and term variation analysis for COVID-19 and SARS-CoV-2. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020.

41. Manning, C.D., and Schutze, H. (1999). Foundations of Statistical Natural Language Processing (MIT Press).

42. Leaman, R., Wei, C.H., Allot, A., and Lu, Z. (2020). Ten tips for a text-mining-ready article: how to improve automated discoverability and interpretability. PLoS Biol. *18*, e3000716. https://doi.org/10.1371/journal.pbio.3000716.

43. Wei, C.H., Allot, A., Leaman, R., and Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. *47*, W587–W593. https://doi.org/10.1093/nar/gkz389.

44. Islamaj, R., Yeganova, L., Kim, W., Xie, N., Wilbur, W.J., and Lu, Z. (2020). PDC - a probabilistic distributional clustering algorithm: a case study on suicide articles in PubMed. AMIA Jt Summits Transl. Sci. Proc. *2020*, 259–268.

45. Greenhalgh, T. (1997). How to read a paper. Getting your bearings (deciding what the paper is about). BMJ *315*, 243–246. https://doi.org/10.1136/bmj.315.7102.243.

46. Chan, E.Y.S., Cheng, D., and Martin, J. (2021). Impact of COVID-19 on excess mortality, life expectancy, and years of life lost in the United States. PLoS One *16*, e0256835. https://doi.org/10.1371/journal.pone.0256835.

47. Fu, D., Chen, M., Sala, F., Hooper, S., Fatahalian, K., and Re, C. (2020). Fast and three-rious: speeding up weak supervision with triplet methods. Proceedings of the 37th International Conference on Machine Learning *119*, 3280–3291.

48. Wolpert, D.H. (1992). Stacked generalization. Neural Network. *5*, 241–259.

49. Wallis, S. (2013). Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. J. Quant. Ling. *20*, 178–208. https://doi.org/10.1080/09296174.2013.799918.

50. Allot, A., Lee, K., Chen, Q., Luo, L., and Lu, Z. (2021). LitSuggest: a web-based system for literature recommendation and curation using machine learning. Nucleic Acids Res. *49*, W352–W358. https://doi.org/10.1093/nar/gkab326.