Long Covid: A Comprehensive Collection of Articles on Post-COVID Conditions

Robert Leaman, Ph.D., Qingyu Chen, Ph.D., Alexis Allot, Ph.D.,
Rezarta Islamaj, Ph.D., Zhiyong Lu, Ph.D.
NCBI/NLM/NIH, Bethesda, Maryland, USA

A substantial percentage of COVID-19 survivors experience debilitating symptoms long after the acute phase has resolved (1). The various names used in the literature for this complex condition include Long Covid - the name typically preferred by patients/advocates - but also post-COVID conditions, post-acute sequelae of SARS-CoV-2, and many others. This considerable variation makes locating articles on Long Covid challenging: precise queries return limited results, but broader queries suffer from low accuracy. We therefore created a comprehensive, searchable collection of Long Covid articles as an extension to LitCovid, a widely used literature hub with nearly 180,000 articles specific to COVID-19 (2). We hope that the Long Covid collection, updated weekly, will help researchers and healthcare professionals keep up with the latest research. The Long Covid collection is searchable online at the LitCovid portal: https://www.ncbi.nlm.nih.gov/research/coronavirus/docsum?text=e_condition:LongCovid

Methods

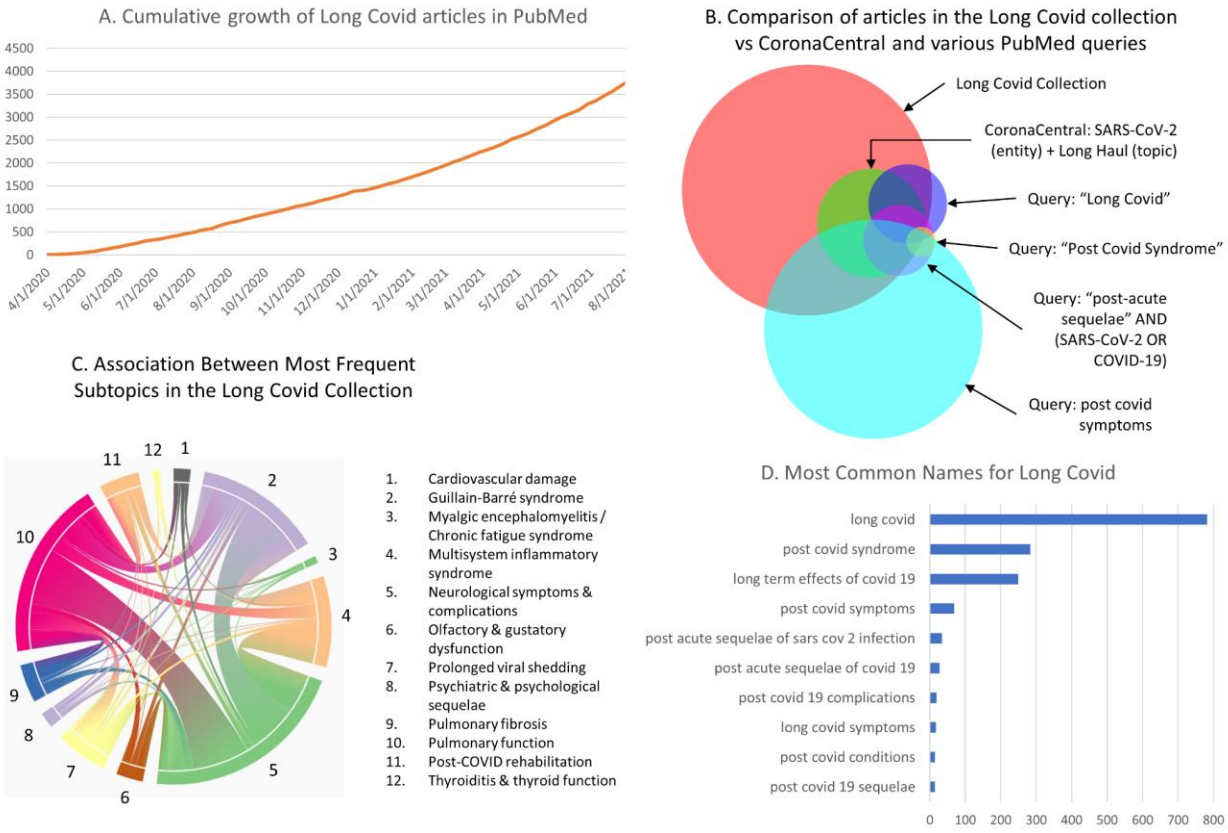
We created the Long Covid collection using a human-in-the-loop machine learning process, allowing all articles relevant to Long Covid to be identified without manually reviewing every article. Following early definitions of Long Covid, we defined an article to be relevant if it considers symptoms or complications beyond 4 weeks from the onset of COVID-19 symptoms.

Our process combines multiple relevance signals into a single probabilistic prediction, which is then used to choose new articles for manual annotation. The automated relevance signals are then updated using the newly-annotated articles, iteratively improving the predictions. Our relevance signals include: 1. predictions from the LitSuggest web-based literature curation tool (3), 2. a purpose-built pattern-based named entity recognition (NER) system to identify mentions of Long Covid, 3. entity annotations from PubTator, 4. MeSH terms, and 5. predictions from CoronaCentral (4). We combine these relevance signals using triplet data programming (5), which estimates the accuracy of each input from the agreement rates between input pairs. Our extension of triplet data programming allows the full range of probabilities, not only binary predictions, and improves the reliability of the accuracy estimates by ignoring input pairs where their agreement may be due to random chance.

Results

As of early October 2021, the Long Covid collection contains 4,643 articles. Figure 1a shows the number of articles on Long Covid over time. Figure 1b compares the set of articles relevant to

Long Covid to the articles returned by querying PubMed for various Long Covid terms and the articles in CoronaCentral annotated with the entity "SARS-CoV-2" and topic "Long Haul." Note the low recall of precise queries (e.g. "Long Covid") and low precision of broad queries (e.g. post covid symptoms), confirming the difficulty of locating Long Covid articles by query alone. Figure 1c shows the number of articles within 12 most frequent subtopics identified by the PDC clustering algorithm, with pairwise associations. Figure 1d shows the Long Covid terms found most frequently in the collection by the NER system.



Acknowledgments

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

References

1. Nalbandian A, Sehgal K, Gupta A, Madhavan MV, McGroder C, Stevens JS, et al. Post-acute COVID-19 syndrome. Nat Med. 2021 Apr;27(4):601–15.
2. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Research. 2021 Jan 8;49(D1):D1534–40.

3. Allot A, Lee K, Chen Q, Luo L, Lu Z. LitSuggest: a web-based system for literature recommendation and curation using machine learning. Nucleic Acids Research. 2021 Jul 2;49(W1):W352–8.
4. Lever J, Altman RB. Analyzing the vast coronavirus literature with CoronaCentral. PNAS. 2021 Jun 8;118(23).
5. Fu D, Chen M, Sala F, Hooper S, Fatahalian K, Re C. Fast and Three-rious: Speeding Up Weak Supervision with Triplet Methods. In: International Conference on Machine Learning. PMLR; 2020 p. 3280–91.