Open Forum Infectious Diseases

MAJOR ARTICLE

Machine Learning-based Analysis of Publications Funded by the National Institutes of Health's Initial COVID-19 Pandemic Response

Anirudha S. Chandrabhatla,[1,a] Adishesh K. Narahari,[2,a] Taylor M. Horgan,[1] Paranjay D. Patel,[3] Jeffrey M. Sturek,[1,4] Claire L. Davis,[1,4] Patrick E. H. Jackson,[1,5] and Taison D. Bell[1,4,5]

[1]School of Medicine, University of Virginia, Charlottesville, Virginia, USA, [2]Division of Cardiothoracic Surgery, University of Virginia School of Medicine, Charlottesville, Virginia, USA, [3]Department of Cardiovascular Surgery, Houston Methodist Hospital, Houston, Texas, USA, [4]Division Of Pulmonary and Critical Care Medicine, University of Virginia, Charlottesville, Virginia, USA, and [5]Division of Infectious Diseases and International Health, University of Virginia, Charlottesville, Virginia, USA

Background. The National Institutes of Health (NIH) mobilized more than $4 billion in extramural funding for the COVID-19 pandemic. Assessing the research output from this effort is crucial to understanding how the scientific community leveraged federal funding and responded to this public health crisis.

Methods. NIH-funded COVID-19 grants awarded between January 2020 and December 2021 were identified from NIH Research Portfolio Online Reporting Tools Expenditures and Results using the "COVID-19 Response" filter. PubMed identifications of publications under these grants were collected and the NIH *iCite* tool was used to determine citation counts and focus (eg, clinical, animal). *iCite* and the NIH's *LitCOVID* database were used to identify publications directly related to COVID-19. Publication titles and Medical Subject Heading terms were used as inputs to a machine learning–based model built to identify common topics/themes within the publications.

Results and Conclusions. We evaluated 2401 grants that resulted in 14 654 publications. The majority of these papers were published in peer-reviewed journals, though 483 were published to preprint servers. In total, 2764 (19%) papers were directly related to COVID-19 and generated 252 029 citations. These papers were mostly clinically focused (62%), followed by cell/molecular (32%), and animal focused (6%). Roughly 60% of preprint publications were cell/molecular-focused, compared with 26% of nonpreprint publications. The machine learning–based model identified the top 3 research topics to be clinical trials and outcomes research (8.5% of papers), coronavirus-related heart and lung damage (7.3%), and COVID-19 transmission/epidemiology (7.2%). This study provides key insights regarding how researchers leveraged federal funding to study the COVID-19 pandemic during its initial phase.

Keywords. COVID-19; machine learning; natural language processing; NIH Funding; SARS-CoV-2.

The National Institutes of Health (NIH) is the main public funding source for biomedical research in the United States. Through a combination of regular and special appropriations such as the *Coronavirus Aid, Relief, and Economic Security* act, the NIH played a major role in funding research to advance knowledge regarding SARS-CoV-2 and the COVID-19 pandemic [1]. Achieving this goal required researchers to perform and publish high-quality science that contributed to this rapidly changing field. Though much public attention was given to clinical research regarding therapeutic and vaccine development, other important research related to SARS-CoV-2 transmission, epidemiology-based testing, and social determinants of health were needed to thoroughly study the pandemic and inform our response.

To date, there has not been a comprehensive evaluation of COVID-related research funded by the NIH. Specifically, no one has studied where NIH-funded COVID research has been published and what major topics have been highlighted. Machine-learning based "topic modeling" is suited to conduct this type of analysis. Topic modeling is a type of natural language processing used to analyze large collections of text to identify major topics/themes. For example, topic modeling on publications from various fields has been used to analyze the research landscape and possibilities for future development [2–8]. More recently, topic modeling has been used to analyze

Received 18 February 2024; editorial decision 10 March 2024; accepted 14 March 2024; published online 24 April 2024

[a]These authors contributed equally to this work.

Correspondence: Taison Bell, MD, MBA, Department of Medicine, Division of Pulmonary and Critical Care, University of Virginia Health Sciences Center, 1215 Lee St, Charlottesville, VA 22903 (tdb4c@uvahealth.org).

Open Forum Infectious Diseases®

https://doi.org/10.1093/ofid/ofae156

online activity to assess public opinions and perceptions regarding the COVID-19 pandemic [9–14].

We took a 2-step approach to quantify and comprehensively assess the NIH-funded research conducted in response to the COVID-19 pandemic. First, we assessed where COVID research was being published, how many citations were being generated, and how publications were split among basic, translational, and clinical research. We next trained a topic model to assess what research areas were investigated by publications resulting from NIH-funded COVID-19 grants.

METHODS

Grant and Publication Information

The National Institutes of Health's Research Portfolio Online Reporting Tools Expenditures and Results (RePORTER) database was queried to collect grant information. NIH RePORTER contains grant information such as funding, supported publications, and relevant patent information. Per guidelines on the NIH website, COVID-specific grants (ie, grants awarded to study COVID-19 via special COVID-related federal appropriations) between January 2020 and December 2021 were identified using the "NIH COVID-19 Response" filter. A Python script using Phantom JS and BeautifulSoup was used to collect the PubMed Identification and journal of each publication linked to NIH COVID funding. We used the NIH *iCite* tool to determine the citation count and paper focus (eg, clinical, animal, molecular/cell) for each publication and a combination of *iCite* and the NIH's *LitCOVID* database [15–17] to determine which publications were directly related to COVID-19. The *iCite* tool is maintained by the NIH's Office of Portfolio Analysis and houses a "COVID-19 Portfolio" of more than 370 000 publications that is deemed the "…NIH's comprehensive, expert-curated source for publications and preprints related to either COVID-19 or the novel coronavirus SARS-CoV-2" [18]. Similarly, the *LitCOVID* database uses machine learning-based classification to identify COVID-related publications (ie, publications that specifically investigate COVID-19 and SARS-CoV-2) and has 380 000 identified papers as of November 2023. Preprints that were subsequently published in peer-reviewed journals were not double counted in the analysis.

Topic Modeling

Machine learning (ML)-based topic modeling was conducted using *BERTopic*, which is a Python-based natural language processing algorithm that clusters text into human-interpretable topics by using numerical representations of words. These numerical representations are similar for words and phrases with close meaning, thereby allowing the algorithm to identity commonalities within text [19]. Publication titles and Medical Subject Heading (MeSH) terms were used as inputs to the model (performed by author A.S.C.). K-means clustering was used to categorize each publication into 20 topics across basic/translational and clinical research [20]. The *BERTopic* model outputs a list of words that are most representative for each topic and this list was used to name the topics (performed by authors A.K.N. and T.M.H.). To assess the model's performance in assigning topics, we randomly selected 150 papers for manual topic validation, ensuring that each topic had at least 5 papers in the validation set. Validation was conducted by reading the abstracts of papers and assessing concordance between the model-assigned topic and the primary topic of the publication. The *BioC* PubMed application programming interface was used to retrieve paper titles, abstracts, and publication year/month [21]. Paper titles and abstracts were used as inputs into the *BERTopic* model. All code was executed on the University of Virginia's Rivanna high-performance computing core.

RESULTS

Journal Analyses

We identified 2401 COVID-19-related grants that resulted in 14 654 unique publications published across 2621 unique journals. We identified the top 50 journals with the most publications (Figure 1*A*). These 50 journals accounted for 4021 (28%) of the publications. The preprint archive medRxiv had the most overall publications with 208. bioRxiv, another preprint, was also in the top 10 with 146 unique publications. The first preprint paper in the data analyzed here was published in March 2020. Between March 2020 and November 2021, there was an overall decline in the percentage of papers each month that were published in preprint servers (Supplementary Figure 1).

High-profile journals and their affiliates were also represented in the top 50 overall list. To better assess the role high-profile journals played in publishing research from NIH-funded COVID grants, we analyzed the number of publications in 6 journals and their affiliates: *Journal of the American Medical Association*, *New England Journal of Medicine*, *Nature*, *Cell*, *Lancet*, and *Science*. Overall, there were 1873 (13% of total) papers published in these journals, with the most in *Nature* and its affiliates such as *Nature Communications* (935) (Figure 1*B*).

Research Patterns

Of the 14 654 publications, 2764 (19%) were identified as directly related to COVID-19 per the NIH *iCite* tool and PubMed *LitCOVID* database. Of the papers directly related to COVID-19, the proportion published in high-impact journals and their affiliates decreased significantly over time ($P = .02$; Supplementary Figure 2). Overall, 62% of papers directly related to COVID-19 were clinically focused, followed by 32% cell/molecular, and 6% animal-focused. This distribution differed between preprint and nonpreprint publications as seen in Figure 2, with a relatively higher percentage of cell/molecular papers in the preprint subgroup. The 2764 papers generated

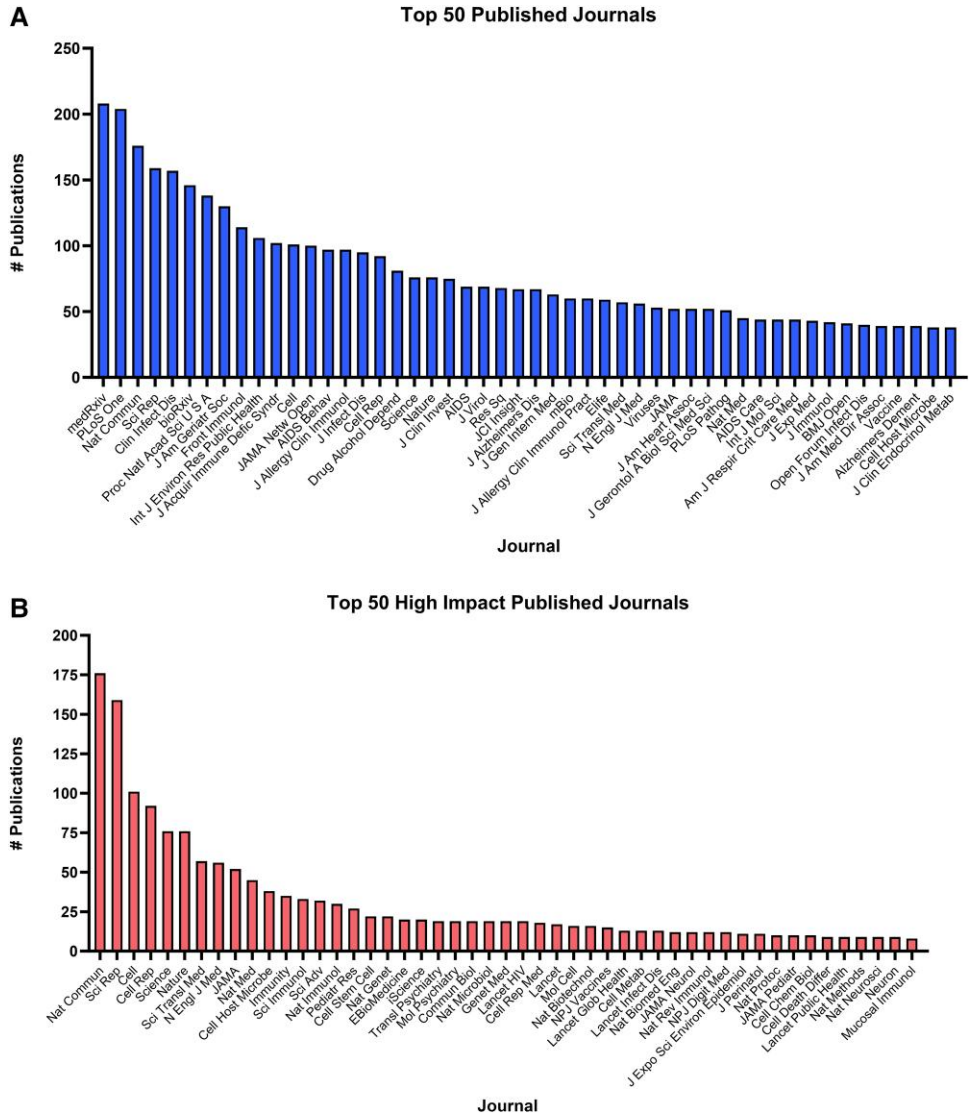

Figure 1. Distribution of papers across journals. (*A*) The top 50 overall published journals accounted for 28% of all publications analyzed here. The pre-print journal medRxiv was the most published overall, with 208 unique publications. (*B*) Publications in six "high profile" journals and their affiliates accounted for 13% of all publications. 935 publications were in nature and its affiliates.

252 029 citations (52% of citations from clinically focused papers, 39% molecular/cellular, 8% animal). Average yearly citation rate was significantly higher for animal compared with clinical studies (58 vs 33, $P = .003$) and cell/molecular compared with human studies (43 vs 33, $P = .03$). There was no difference in animal compared with cell/molecular studies (58 vs 43, $P = .09$). The top 10 most cited publications are seen in Table 1. Interestingly, of these 10 studies, only 2 were clinical trials. Publication ramp-up analysis revealed that cell/molecular papers were published at the fastest rate, followed by animal and clinical papers (Figure 3).

MeSH Term and ML-based Topic Analysis

Topic modeling was conducted to assess the top 20 areas of research that were explored by the COVID-related papers. The top 5 topics by number of publications were: clinical trials and outcomes research (234 publications, 8.5%), coronavirus-related heart and lung damage (202, 7.3%), COVID-19 transmission/epidemiology (199, 7.2%), inflammation and systemic manifestations of COVID-19 (194, 7.0%), and vulnerable/disadvantaged populations (193, 7.0%) (Figure 4*A*). A 2-dimensional representation of the topics is seen in Figure 4*B*. The topic distribution of preprint versus nonpreprint publications is seen in Supplementary Table 1. The largest difference was seen for the "ACE and spike protein" topic, which made up 12.06% of preprint publications, but 5.16% of nonpreprint publications. The next largest difference was for the "Vulnerable and disadvantaged populations" topic, which accounted for 1.32% of preprint publications, but 8.10% of nonpreprint publications. Manual verification of 150 randomly

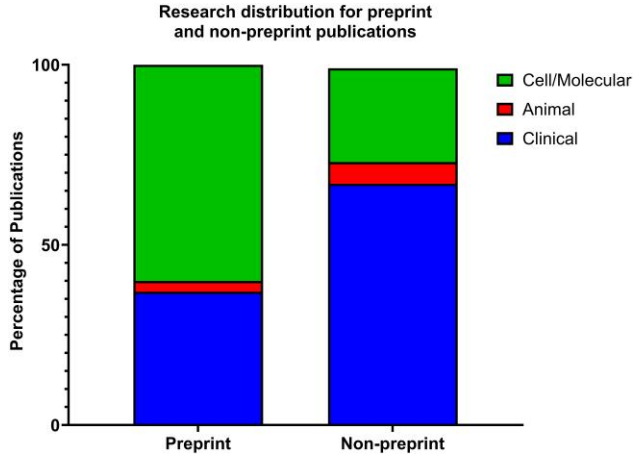

Figure 2. Research distribution across preprint and non-preprint journals. Preprint journals had a relatively higher proportion of cell/molecular studies compared to non-preprint journals.

selected publications revealed an overall model accuracy of 90% in assigning topics (Supplementary File 1). The top 20 MeSH terms in preprint publications accounted for 50% of all MeSH terms in the preprint cohort. Of these MeSH terms, ~50% were related to COVID-19 immunology and epidemiology (Figure 5).

DISCUSSION

The COVID-19 pandemic presented a unique, generational challenge to the scientific community. Researchers across public and private institutions needed to quickly mobilize to investigate the behavior of the SARS-CoV-2 virus, develop therapies and vaccines, and build the foundation for a future in which the virus could be managed. Here, we used unsupervised machine learning to analyze more than 14 000 publications that were published under 2401 COVID-19–related NIH grants. Surprisingly, the majority (~83%) of papers in this cohort were not directly related to COVID-19. We identified that the research community broadly studied the pandemic, investigating not only basic, translational, and clinical science, but also answering questions regarding health disparities and vaccine hesitancy. We also report a difference in the types of research published in traditional journals versus preprint servers.

Preprint servers emerged as relatively new vectors of scientific communication that were particularly useful in publishing COVID research. The COVID-19 pandemic highlighted the advantage preprints have in quickly disseminating key scientific information, as preprint servers such as *bioRxiv* and *medRxiv* were both in the top 5 most published journals. Interestingly, 60% of preprint publications were cell-/molecular-focused studies, compared with only 26% of nonpreprint publications. This is also reflected in preprint publications' most common MeSH terms, which had strong representation from basic science topics such as immunology, and in the topic distribution of preprint compared to nonpreprint publications, with preprints having a higher percentage of papers in the "ACE and spike protein," "Wastewater and genomic testing," and "mRNA technology" topics. During the pandemic, this fast dissemination of basic science in preprint servers may have supported hypothesis generation and preliminary validation for groups to augment their research. The decision to post a paper to a preprint server can be complex and include factors such as speed of dissemination, gathering feedback from other groups, or publishing negative results [22]. There are likely many opinions regarding which research topics are better suited for initial publication to preprint servers, but the higher proportion of cell/molecular studies in preprint servers indicates that researchers may have wanted to expedite access to highly valuable data early in the pandemic to foster further scientific inquiry and collaboration at the basic science level in areas such as SARS-CoV-2 transmission and vaccine development. There might also be an element of authors initially submitting to preprints if they believed their work could be subject to a long peer-review process because, perhaps, of a shortage of reviewers in their target journals during the pandemic. The research community must continue to cautiously interpret results published in preprint servers as they lack the rigorous peer review process offered by traditional journals. Traditional journals pivoted their priorities to facilitate publication of COVID-related science and continued to play a major role in scientific communication.

Per grant, the grants analyzed here accounted for roughly 6 papers overall and roughly 1 COVID-specific paper. Retrospective analyses of NIH grants over the past 20 years have reported an average of 7 to 17 publications per grant, with a mean time to initial publication of 15 months from the grant's start date [23, 24]. The 2401 grants analyzed here were within their first 2 years of funding, yet still published 6 papers per grant, indicating scientific productivity above the historical average. Interestingly, only 17% of papers in our analysis were specifically related to COVID-19. The "non-COVID" papers were either studying COVID-adjacent topics (eg, general virology, vaccine distribution) or papers published by research groups that happened to also have individuals awarded COVID-related grant funding. It should be noted, however, that COVID-adjacent papers, though they may not have directly studied the pandemic or SARS-CoV-2, still contributed to our collective work to understand and control the pandemic. Even still, this amounts to an average of 100 publications per month in the analyzed period, further indicating robust scientific productivity. This research output lead to higher citations per year for cell/molecular and animal studies compared with clinically focused publications. This is likely the result of a relatively smaller number of cell/molecular and

Table 1. Top 10 Most Cited COVID-related Publications

Publication Title	Clinical Trial?	Journal	Number of Citations
Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine	Yes	*New England Journal of Medicine*	6014
Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area	No	*Journal of the American Medical Association*	5848
Remdesivir for the Treatment of COVID-19—Final Report	Yes	*New England Journal of Medicine*	4552
Integrated analysis of multimodal single-cell data	No	*Cell*	3070
The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application	No	*Annals of Internal Medicine*	3018
Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19	No	*Cell*	2741
A SARS-CoV-2 protein interaction map reveals targets for drug repurposing	No	*Nature*	2739
Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus	No	*Journal of Virology*	2665
The proximal origin of SARS-CoV-2	No	*Nature Medicine*	2523
Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals	No	*Cell*	2457

Three of the publications are clinical trials and there are 6 unique journals represented here.

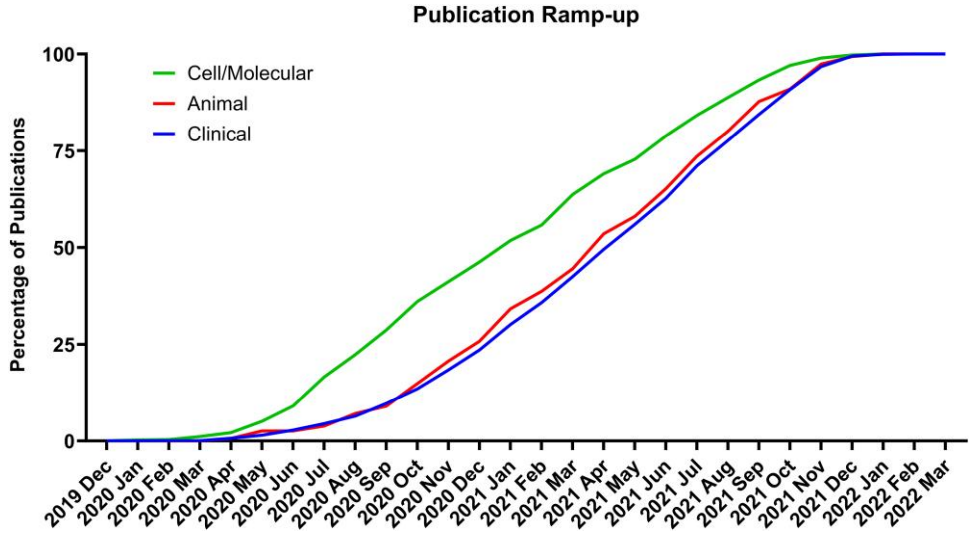

Figure 3. Publication ramp-up. Cell/molecular-oriented studies were published at the fastest rate, followed by animal and then clinically-oriented papers, following the basic to translational to clinical research paradigm.

animal studies and a heightened focus in the scientific community about discovering basic mechanisms of viral infectivity, transmission, and immunity.

The top 20 research areas identified by the ML-based model demonstrated the broad range of topics that the scientific community studied. The top 5 topics accounted for roughly 37% of the 2401 COVID-related publications and included research across the basic-translational-clinical spectrum. The topic analysis revealed that there was not only a focus on improving our understanding of SARS-CoV-2 infectivity and therapeutics, but also on developing a better understanding about how the pandemic impacted mental health and vulnerable/disadvantaged populations. Interestingly, there were 3 separate topics that covered different aspects of vaccinations: "Vaccination and antibody response," "Vaccine distribution, promotion and hesitancy," and "mRNA technology." The model further reveals that future studies regarding the pandemic could bolster relatively less represented research domains such as "long COVID" and future pandemic preparedness. Interestingly, the topics revealed in our model mirror what other groups have found in different topic model implementations to study research output [7, 25, 26]. Our topic model performance of 90% accuracy also compares favorably with other accuracies of BERTopic models in the literature [27, 28].

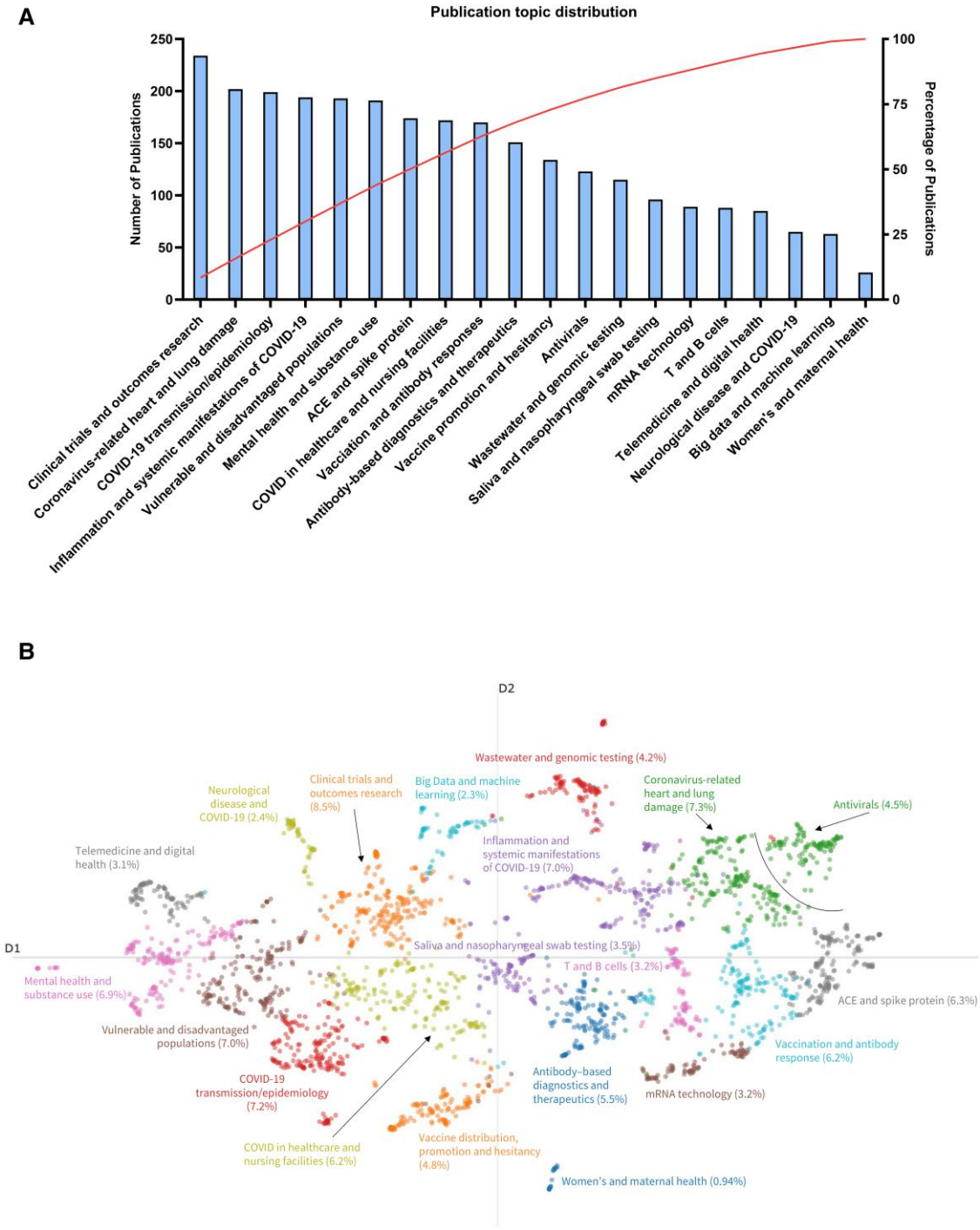

Figure 4. Results of ML-based topic modeling. (*A*) The 20 topics that the unsupervised model discovered in the 2764 COVID-related papers. The first six topics accounted for roughly 50% of the publications. (*B*) Spatial distribution of the 20 topics.

Our study had a few limitations. First, we only analyzed grant funding from the NIH and grants funded by other federal agencies such as the Department of Defense or private institutions were not analyzed. Second, the *iCite* tool used to assess citations only documents citations from studies listed on PubMed and therefore underestimates the total citation count. Third, we used the *iCite* and *LitCOVID* databases to determine "COVID-related" versus "non–COVID-related" publications and may have misclassified some publications. However, both tools were created and are maintained by the NIH and the percentage of misclassified publications is likely negligible.

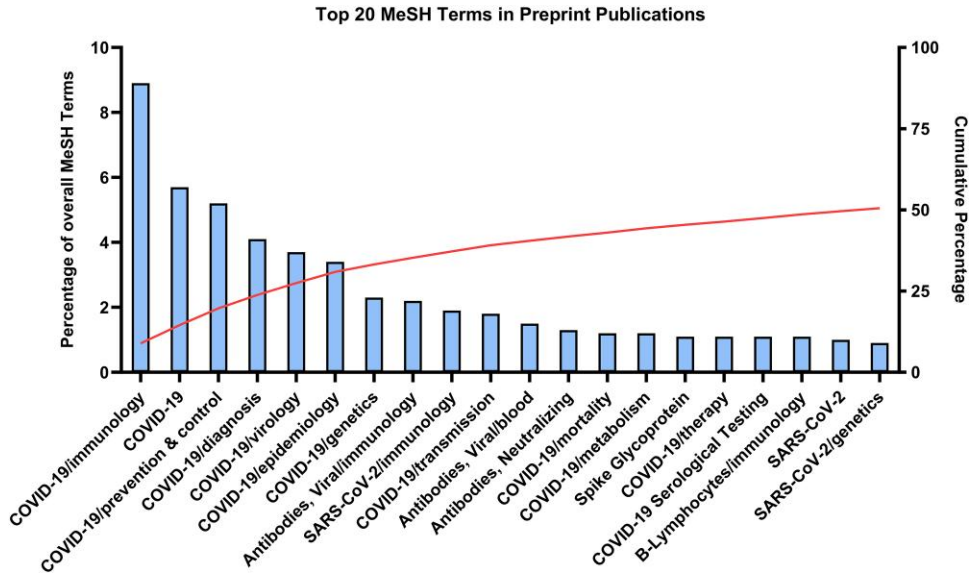

Figure 5. Top 20 MeSH terms in pre-print publications. The top 20 MeSH terms accounted for 50% of MeSH terms associated with the 483 pre-print publications. This analysis shows the focus of these publications on COVID-19 immunology and epidemiology.

CONCLUSIONS

During the COVID-19 pandemic, the NIH rapidly mobilized extramural funding to enable significant research output from public and private institutions. By conducting ML-based topic modeling on more than 2700 papers from 2401 unique COVID-19 related NIH grants, we identified broad themes within the research that spanned areas such as virology and vaccine development, but also began investigating health disparities and inequities. Preprint servers were important mediators for distributing science across the basic-translational-clinical spectrum, though COVID-related papers in preprints were predominantly cell-/molecular-focused. Our study provides insight on how the scientific community used NIH funding to study the COVID-19 pandemic in its initial stages. Future studies may look into results from COVID-adjacent research, deep-dive into specific topics identified here to understand subtopics of focus or assess the output of COVID-related research funded by non-NIH sources (eg, Department of Defense, private institutions).

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Acknowledgments

Patient consent. This study did not require patient consent.

Financial support. This work was supported by National Institutes of Health (F30CA236370) to A. K. N.

Potential conflicts of interest. All authors: No reported conflicts.

References

1. Courtney J. H.R.748—116th Congress (2019–2020): CARES Act. 2020. Available at: https://www.congress.gov/bill/116th-congress/house-bill/748/text.
2. Feng C, Tian C, Huang L, Chen H, Feng Y, Chang S. A bibliometric analysis of the landscape of parathyroid carcinoma research based on the PubMed (2000–2021). Front Oncol **2022**; 12:824201.
3. Zhang Z, Yao L, Wang W, Jiang B, Xia F, Li X. A bibliometric analysis of 34,692 publications on thyroid cancer by machine learning: how much has been done in the past three decades? Front Oncol **2021**; 11:673733.
4. Li C, Liu Z, Shi R. A bibliometric analysis of 14,822 researches on myocardial reperfusion injury by machine learning. Int J Environ Res Public Health **2021**; 18:8231.
5. Feng C, Wu Y, Gao L, Guo X, Wang Z, Xing B. Publication landscape analysis on gliomas: how much has been done in the past 25 years? Front Oncol **2020**; 9:1463.
6. Hai Ha G, Thanh-Phan H, Vu GT, et al. Fertility desire in HIV/AIDS research during 1992–2019: a systematic text mining of global literature. AIDS Rev **2021**; 22:213–20.
7. Tran BX, Phan HT, Nguyen QN, et al. Pre-exposure prophylaxis in HIV research: a latent Dirichlet allocation analysis (GAPRESEARCH). AIDS Rev **2021**; 22:103–11.
8. Baghaei Lakeh A, Ghaffarzadegan N. Global trends and regional variations in studies of HIV/AIDS. Sci Rep **2017**; 7:4170.
9. Cao G, Shen L, Evans R, et al. Analysis of social media data for public emotion on the Wuhan lockdown event during the COVID-19 pandemic. Comput Methods Programs Biomed **2021**; 212:106468.
10. Ntompras C, Drosatos G, Kaldoudi E. A high-resolution temporal and geospatial content analysis of Twitter posts related to the COVID-19 pandemic. J Comput Soc Sci **2021**; 5:1–43.
11. Cotfas L-A, Delcea C, Gherai R. COVID-19 vaccine hesitancy in the month following the start of the vaccination process. Int J Environ Res Public Health **2021**; 18:10438.
12. Hampshire A, Hellyer PJ, Trender W, Chamberlain SR. Insights into the impact on daily life of the COVID-19 pandemic and effective coping strategies from free-text analysis of people's collective experiences. Interface Focus **2021**; 11:20210051.
13. Zheng C, Xue J, Sun Y, Zhu T. Public opinions and concerns regarding the Canadian Prime Minister's daily COVID-19 briefing: longitudinal study of YouTube comments using machine learning techniques. J Med Internet Res **2021**; 23:e23957.
14. Patel J, Desai H, Okhowat A. The role of the Canadian media during the initial response to the COVID-19 pandemic: a topic modelling approach using Canadian broadcasting corporation news articles. JMIR Infodemiology **2021**; 1:e25242.
15. Chen Q, Allot A, Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Res **2021**; 49:D1534–40.
16. Chen Q, Allot A, Leaman R, et al. LitCovid in 2022: an information resource for the COVID-19 literature. Nucleic Acids Res **2023**; 51:D1512–8.
17. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. Nature **2020**; 579:193.

18. NIH Office of Portfolio Analysis. iCite COVID-19 Portfolio. Available at: https://icite.od.nih.gov/covid19/search/#search:searchId=6556d3e13089f55f52556767.
19. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. 2022; Available at: http://arxiv.org/abs/2203.05794.
20. Likas A, Vlassis NJ, Verbeek J. The global k-means clustering algorithm. Pattern Recognit **2003**; 36:451–61.
21. Comeau DC, Wei C-H, Islamaj Doğan R, Lu Z. PMC text mining subset in BioC: about three million full-text articles and growing. Bioinformatics **2019**; 35:3533–5.
22. Funk K. NIH preprint pilot expands to include preprints across NIH-funded research—NIH Extramural Nexus. 2023; Available at: https://nexus.od.nih.gov/all/2023/02/08/nih-preprint-pilot-expands-to-include-preprints-across-nih-funded-research/.
23. Druss BG, Marcus SC. Tracking publication outcomes of National Institutes of Health grants. Am J Med **2005**; 118:658–63.
24. Riley WT, Bibb K, Hargrave S, Fearon P. Publication rates from biomedical and behavioral and social science R01s funded by the National Institutes of Health. PLoS One **2020**; 15:e0242271.
25. Älgå A, Eriksson O, Nordberg M. Analysis of scientific publications during the early phase of the COVID-19 pandemic: topic modeling study. J Med Internet Res **2020**; 22:e21559.
26. Li D, Wang Z, Wang L, et al. A text-mining framework for supporting systematic reviews. Am J Inf Manag **2016**; 1:1–9.
27. da Silva RP, Pollettini JT, Pazin Filho A. Unsupervised natural language processing in the identification of patients with suspected COVID-19 infection. Cad Saude Publica **2023**;39:e00243722.
28. Scarpino I, Zucco C, Vallelunga R, Luzza F, Cannataro M. Investigating topic modeling techniques to extract meaningful insights in Italian long COVID narration. BioTech (Basel) **2022**; 11:41.