Annals of Library and Information Studies
Vol. 70, March 2023, pp. 41-51
DOI: 10.56042/alis.v70i1.71939

# Topic modelling-based analysis of COVID-19 vaccine articles published in the preprint server MedRxiv

Nishad Deshpande[a], Virendra Ligade[b], Shabib-Ahmed Shaikh[c] and Alok Khode[d]

[a,c,d]CSIR-Unit for Research and Development of Information Products, Pune, Maharashtra, India, Email: nishad@urdip.res.in, shabib@urdip.res.in, alok@urdip.res.in

[b]Department of Pharmacy Management, Manipal College of Pharmaceutical Science, MAHE, Manipal, Karnataka, Email: virendra.sl@manipal.edu

*Received: 08 March 2023*; *accepted: 28 March 2023*

Two thousand one hundred and ninety-eight research publications on COVID-19 vaccines in MedRxiv preprint repository during January 01, 2020 and December 31, 2021 were analyzed for topic modelling with unsupervised inference method. Latent Dirichlet Allocation (LDA) method was used to investigate the thematic structure of the preprints. It was observed that the published articles were related to either clinical trials or patient responses to vaccine or modelling for various applications such as infection transmission, vaccine allocation, vaccine hesitancy etc.

**Keywords:** COVID-19, Vaccine, Preprints, LDA, Topic modelling

## Introduction

Coronavirus disease-2019 or COVID-19 pandemic as it is generally referred to, has made a negative impact on almost all countries in the world. The pandemic not only spread quickly to all parts of the globe, but also tested the medical and health care facilities of nations boasting of maintaining high quality health care services and best health care standards. Over the decades, strategies that involve vaccination have made a major contribution in the fight against infectious diseases. Vaccination works by triggering defense against a pathogen by imitating its natural interaction with the human immune system. However, acceptance or hesitancy towards vaccination and identification/selection of target pathogen candidates for impending diseases remain major challenges[1]. Despite of these challenges, development and distribution of a safe and effective vaccine against SARS-CoV-2 (COVID-19) generated universal interest as the pandemic was affecting day-to-day activities and impacted all the business.

In the race to get a solution on the coronavirus, researchers across the globe started research work and shared manuscripts providing further leads to other researchers. As article published in reputed peer journals go through peer review process that is time-consuming process, researchers usually publish their preprints in open access repositories such as arXiv.org, MedRxiv, bioRxiv etc. for faster and open dissemination of their work. These preprint repositories do not have peer review. Preprints are not considered as researchers' official publications and generally are considered as precedent of peer-reviewed article[2], however, they are assigned a DOI like journal articles and are indexed in Google Scholar. There are various advantages of preprints which researchers may find beneficial. These advantages can be grouped in three broad areas. The first advantage is the credit attribute that enables citation as preprints facilitate authors to enable or establish ownership for their work by establishing a public record. The second major advantage provided by preprints is the visibility, as the manuscripts in the preprints are open access, thereby making it easy for other researchers to discover, access and cite the preprint. Third advantage is related to the review as preprints can add-on to the traditional peer review by enabling discovery and full text access of the manuscript to greater group of researchers who can contact the research author with recommendations for enhancements that might improve the quality of the work[3]. During the pandemic, preprints increased significantly. This rapid progress has been catalyzed, partly, by the COVID-19-related preprints that have been uploaded during the pandemic period by various researchers, thereby signaling to researchers as well as to the public that publication in the scientific domain can happen at a faster pace and can be open

and accessible to all[4]. However, use of preprints and evolving methods of publication need caution and further scrutiny. Researchers[5] are of the opinion that preprints should not merely describe the data that is available in public domain, rather the preprints should also deliberate on the research methods, and extensively discuss the research presented, and not restrict to displaying few graphs based on lesser amount of data and use it for justification. Benefits of preprints are generally realized through early access to research, but these benefits may be undermined by the release, dissemination, and misuse of unreliable evidence[6]. It has also been suggested[7] that rather than just listing in the reference list, the preprint status should also be mentioned in the text of manuscript appropriately.

*LDA topic modelling*

Topic modelling is an important and powerful data mining techniques that is based on the text mining approach. It has various applications such as discovery of the underlying (latent) data and revealing associations or buried relationship amongst data and text documents. Topic modelling methods are extensively used by the researchers in the natural language processing for discovery of topics and also for semantic mining from unordered documents[8]. Researchers from various fields such as software engineering, political science and medical sciences have applied topic modelling in their research resulting into publication of various scientific articles. Topic modelling can be carried out using various approaches and one of the commonly used approach method used in topic modelling is Latent Dirichlet Allocation (LDA)[8]. Topic models are most suitable for finding and representing discrete data and to discover hidden structures/semantics in huge information. Researchers have proposed various models based on the LDA in topic modelling and LDA Topic modelling has been used in combination with natural language processing in various application such as social media-sentiment analysis and for information retrieval. Topic modelling with LDA can be used to investigate the "hidden" thematic structure of a large amount of text documents quickly and efficiently[9]. When initiated, the LDA algorithm processes set of textual documents and outputs number of topics. The topics are nothing but clusters of words that have repeatedly occurred together in the text corpus and thus LDA model helps in unrevealing topical structure[10]. LDA topic modelling has been applied on textual data harvested from multiple Reddit communities focusing on the COVID-19 vaccine to analyse the public sentiments on vaccines[11]. Latent Dirichlet Allocation for topic modelling has been applied by researcher[12] on abstracts published up to March 2020 from the COVID-19 Open Research Dataset to build eight topic models that yielded various topics (clusters) such as clinical characterization, pathogenesis research, therapeutics research, epidemiological study, virus transmission, vaccines research, virus diagnostics and viral genomics as research hotspots.

As per the literature survey, it was found that not much work has been carried involving topic modelling based on COVID-19 vaccine related research published in preprints server. The present paper attempts to apply LDA for topic modelling on the COVID-19 vaccine articles deposited in the MedRxiv preprint server during the pandemic. Topic modelling was carried out for titles and abstracts of the articles, which were downloaded from preprint server MedRxiv. We used Coherence score as an indicator about the comprehensibility of the topics. We carried out keyword analysis for topic labelling. Some features of the study on COVID-19 vaccine related articles published in the preprint servers could be reflected by the occurrence of keywords. Vaccine design and development, Vaccine response, Vaccine hesitancy and Vaccine effectiveness were top keyword identified from the study.

## Methodology

*Data Collection*

The dataset for topic modelling included title and abstracts of 2,198 articles related to COVID-19 vaccine, which were published in MedRxiv preprint repository between January 01, 2020 and December 31, 2021. The keywords used in searching articles included "COVID vaccine", "SARS-CoV-2 vaccine" and these were searched in title or abstract using "match all words" option. The article bibliographic information including title and abstract were exported into spreadsheet with the help of Zotero (available at https://www.zotero.org/).

*Topic Modelling*

To identify the focus of the articles over the changing time, the 2,198 articles were segregated into quarters for each calendar year and the Topic modelling was carried out for articles published in that respective quarter. The topic modelling was

carried out in Python language, using Jupyter Notebook along with numpy, pandas and Gensim 3.8 libraries of python. In the pre-processing step, the spreadsheet containing title of the articles downloaded from preprint servers viz MedRxiv was read into python with help of openpyxl library. In the next step, bigram and trigram phrases were extracted from the abstract using functions from the python gensim library. As next step of pre-processing, stop words from NLTK library were downloaded and few more stop words were added to remove commonly used words from the bigram and trigrams. The SpaCy python library for advanced natural language processing was used for tokenization and lemmatization of the unique words that formed the corpus in the topic modelling. This pre-processing step is required as it results into extraction of the important data/concepts from into the corpus and imparts precision in topic modelling[13]. The unsupervised clustering of these words from corpus was carried out with the help of LDA MALLET[14] that resulted into multiple clustered topics. The graphs were plotted using python library matplotlib and the word cloud by making use of the seaborn library.

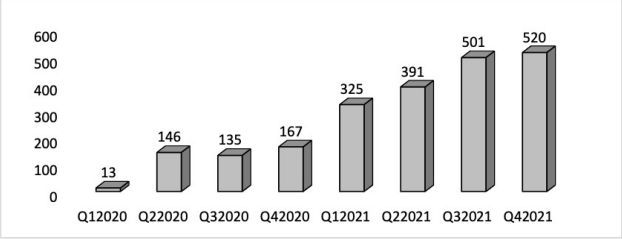


Fig. 1 — Quarter wise publications

## Results

### Analysis

*COVID-19 vaccine related articles published in MedRxiv since January 2020*

A total of 2,198 articles relating to COVID-19 vaccines have been published in MedRxiv during the two years study period from January 1, 2020 to December 31, 2021. Figure 1 highlights the publications in MedRxiv repository for each quarter. It can be observed that there is a steady increase in publication activity over the previous quarter for almost every quarter and almost 6 papers were published daily in the last quarter of 2021.

*Coherence score for 20 Topics*

The number of topic clusters in the topic model can be decided with the aid of expert judgement or with the help of coherence score. Coherence score is used as an indicator about the comprehensibility of the topics[15-18]. To determine optimal number of topics, about 20 topics were generated and resulting coherence scores were recorded.

Twenty clusters were prepared and coherence score for each cluster was plotted. We can use the coherence score in topic modelling to measure how interpretable the topics are. Topics are represented as the top N words with the highest probability of belonging to that topic. The coherence score measures how similar these words are to each other. It measures the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference. From the Figure 2

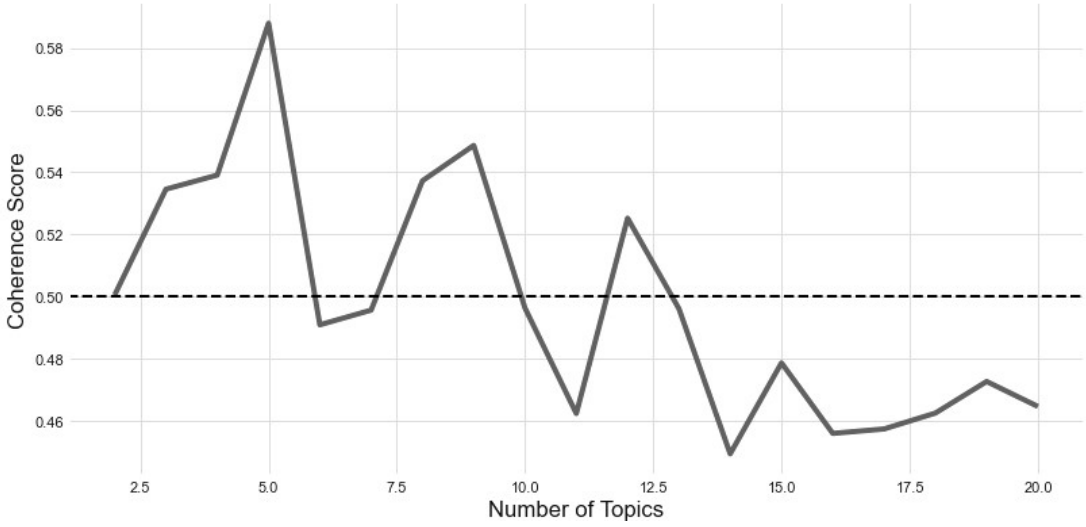


Fig. 2 — Coherence score for 20 Topics

above, it is observed that 5 topics yield highest Coherence score. Accordingly, for every quarter, the articles were clustered in 5 topics. The analysis of the topics for each quarter (*Q1, Q2, Q3* and *Q4*) is discussed below.

*Q1-2020*

In the first quarter of the calendar year 2020 from January 2020 to March 2020, there were only 13 publications available in the preprint repositories. As the available publications were very less in number, non-overlapping topics were not formed in the LDA Mallet model. Hence, publications from this quarter were not considered further for analysis.

*Q2-2020*

During the second quarter of 2020 from the months of April 2020 to June 2020, 146 articles related to COVID-19 vaccines were published in MedRxiv. Figure 3 displays the inter topic distance map for the generated 5 topics for the second quarter of the year 2020. It is observed that there are no overlaps between the topics and the clusters are distinct from each other. The image below presents word cloud for the topics

Table 1 lists the keywords for each generated topic generated, and the labels assigned to the topics.

*Q3-2020*

During the months from July to September 2020, 135 articles related to COVID-19 vaccines were published in MedRxiv. Figure 4 gives the inter topic distance map for the generated 5 topics. It is observed that there are no overlaps between the topics and the clusters are distinct from each other.

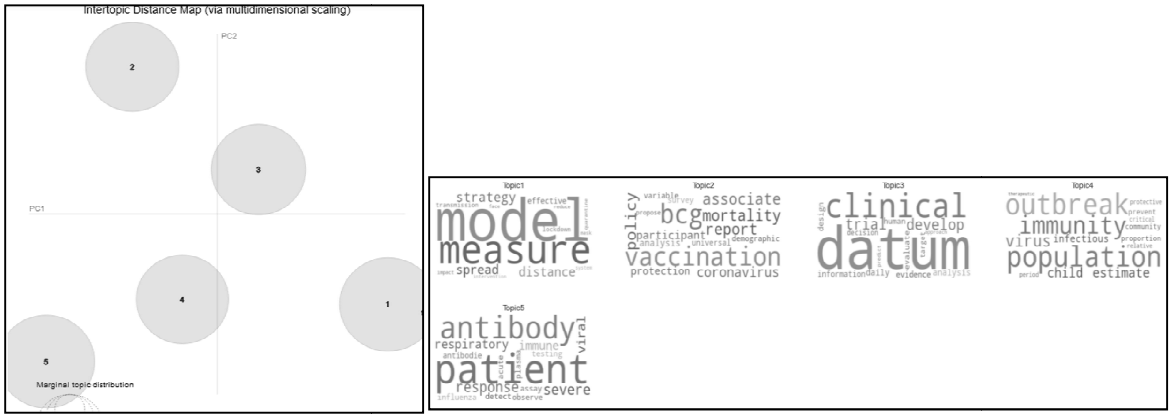


Fig. 3 — Word cloud for the topics generated for preprints from second quarter of year 2020

Table 1–Keywords for each generated topic generated during preprints published in second quarter of year 2020 and the labels assigned to the topics

| | Terms per Topic | Topic Label | Document Count | Percent |
|---|---|---|---|---|
| Topic1 | model, measure, strategy, spread, distance, effective, transmission, lockdown, intervention, reduce, system, impact, quarantine, mask, face | Modelling transmission and spread | 43 | 29.45 |
| Topic2 | vaccination, bcg, mortality, policy, report, associate, coronavirus, participant, protection, analysis, demographic, variable, survey, universal, propose | Vaccination interventions | 29 | 19.86 |
| Topic3 | datum, clinical, develop, trial, analysis, daily, information, evidence, design, evaluate, human, decision, target, approach, predict | Clinical trial | 20 | 13.70 |
| Topic4 | population, outbreak, immunity, virus, estimate, child, infectious, proportion, prevent, community, protective, period, critical, relative, therapeutic | Infection prevalence | 16 | 10.96 |
| Topic5 | patient, antibody, response, severe, respiratory, immune, viral, testing, assay, detect, influenza, antibodie, observe, plasma, acute | Response to vaccines | 38 | 26.03 |
| The dominant topic during second quarter of year 2020 was observed to be Modelling transmission and spread followed by patient response to treatments | | | | |

Table 2 lists the keywords for each generated topic generated and the labels assigned to the topics.

*Q4-2020*

During the months from October to December 2020, 167 articles related to COVID-19 vaccines were published in MedRxiv. Figure 5 below displays the inter topic distance map for the generated 5 topics. It is observed that there are no overlaps between the topics and the clusters are distinct from each other.

Table 3 lists the keywords for each generated topic generated and the labels assigned to the topics.

*Q1-2021*

During the months from January 2021 to March 2021, a total of 325 articles related to COVID-19 vaccines were published in MedRxiv. Figure 6 displays the inter topic distance map for the generated 5 topics. It is observed that there are no overlaps between the topics and the clusters are distinct from each other.

Table 4 lists the keywords for each generated topic generated and the labels assigned to the topics.

*Q2-2021*

During the months from April to June 2021, 391 articles related to COVID-19 vaccines were published in MedRxiv. Figure 7 displays the inter topic distance map for the generated 5 topics. It is observed that there are no overlaps between the topics and the clusters are distinct from each other.

Table 5 lists the keywords for each generated topic generated and the labels assigned to the topics.

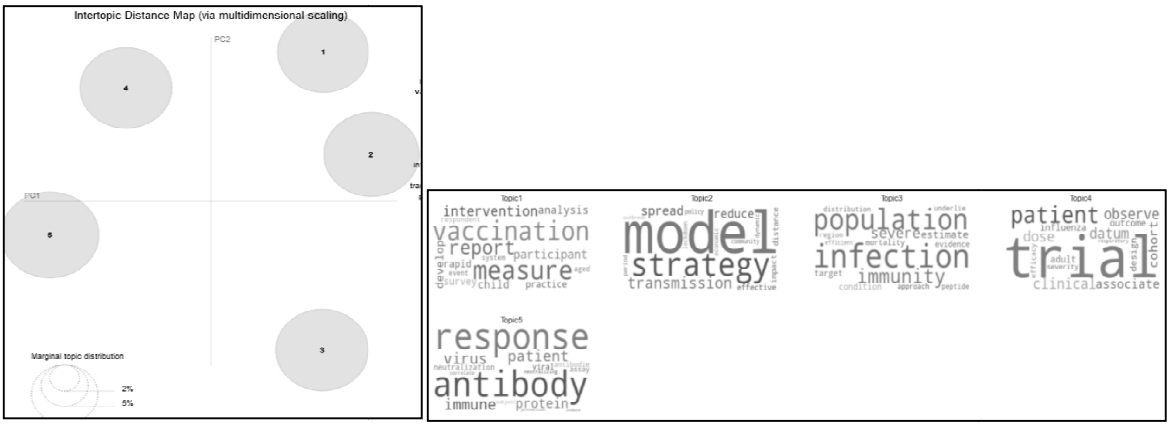


Fig. 4 — Word cloud for the topics generated for preprints from third quarter of year 2020

Table 2 – Keywords for each generated topic generated during preprints published in third quarter of year 2020 and the labels assigned to the topics

| | Terms per Topic | Topic Label | Document count | Percentage |
|---|---|---|---|---|
| Topic1 | vaccination, measure, report, intervention, participant, analysis, develop, child, survey, rapid, practice, respondent, aged, system, event | Vaccination interventions | 18 | 13.33 |
| Topic2 | model, strategy, transmission, spread, reduce, effective, distance, impact, community, policy, economic, lockdown, period, dynamic, outbreak | Models for assessing spread, transmission | 44 | 32.59 |
| Topic3 | infection, population, immunity, severe, estimate, condition, target, evidence, mortality, approach, distribution, region, peptide, underlie, efficient | Infection prevalence | 19 | 14.07 |
| Topic4 | trial, patient, clinical, observe, datum, dose, associate, cohort, influenza, adult, outcome, design, severity, efficacy, respiratory | Clinical trial | 19 | 14.07 |
| Topic5 | response, antibody, patient, virus, immune, protein, neutralization, viral, antibodie, assay, neutralizing, subject, correlate, protection, induce | Response to vaccines | 35 | 25.93 |
| The dominant topic during third quarter of year 2020 was observed to be Modelling transmission and spread followed by patient response to vaccines | | | | |

*Q3-2021*

During the months from July to September 2021, 501 articles related to COVID-19 vaccine were found in MedRxiv. Figure 8 below displays the inter topic distance map for the generated 5 topics. It is observed that there are no overlaps between the topics and the clusters are distinct from each other.

Table 6 lists the keywords for each generated topic generated and the labels assigned to the topics.

*Q4-2021*

During the months from October to December 2021, 521 articles related to COVID-19 vaccine were published in MedRxiv. Figure 9 displays the inter topic distance map for the generated 5 topics. It is observed that there are no overlaps between the topics and the clusters are distinct from each other.

Table 7 lists the keywords for each generated topic generated and the labels assigned to the topics.

**Discussion**

LDA topic modelling revealed a number of topics under which the published articles could be clustered, and the most prominent clusters include various models, patient responses to the vaccines, results from clinical trials, and vaccine hesitancy. Under the topics related to models, we can observe various papers related to vaccine prioritization/allocation or modelling that tries to assess the infection spread etc. The articles under patient response cluster are related

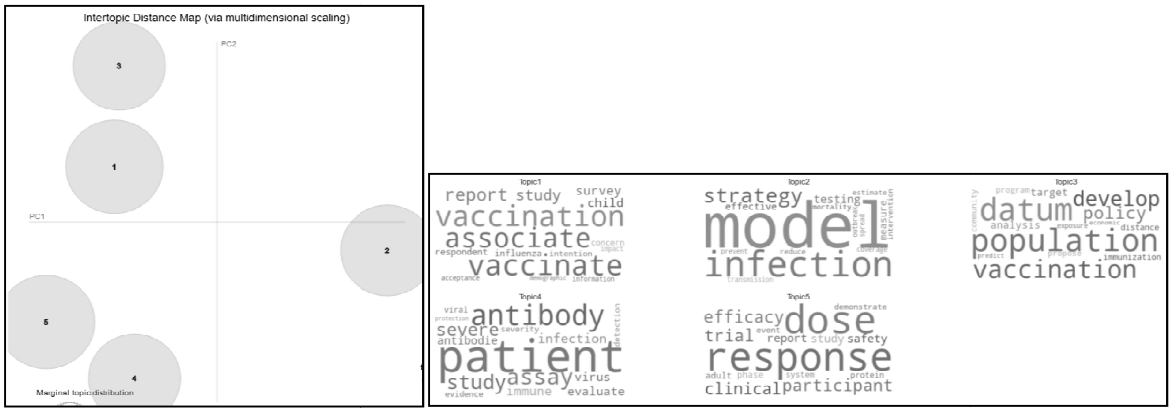


Fig. 5 — Word cloud for the topics generated for preprints from fourth quarter of year 2020

Table 3 – Keywords for each generated topic generated during preprints published in fourth quarter of year 2020 and the labels assigned to the topics

| | Terms per Topic | Topic Labels | Document count | Percentage |
|---|---|---|---|---|
| Topic1 | vaccination, vaccinate, associate, report, study, survey, child, respondent, influenza, intention, concern, acceptance, impact, information, demographic | Vaccination acceptance/concern | 40 | 23.95 |
| Topic2 | model, infection, strategy, testing, effective, measure, mortality, estimate, reduce, intervention, transmission, prevent, coverage, spread, outbreak | Models for assessing spread, transmission | 40 | 23.95 |
| Topic3 | population, datum, vaccination, develop, policy, analysis, target, immunization, program, community, distance, propose, exposure, economic, predict | Vaccination policy | 28 | 16.77 |
| Topic4 | patient, antibody, assay, study, severe, infection, immune, evaluate, virus, antibodie, severity, viral, detection, evidence, protection | Response to vaccines | 38 | 22.75 |
| Topic5 | response, dose, participant, efficacy, trial, clinical, report, study, safety, phase, adult, protein, system, event, demonstrate | Clinical trial | 21 | 12.57 |

The dominant topic during fourth quarter of year 2020 was observed to be Vaccination acceptance/attitude which is a new topic that has emerged in this quarter. Modelling transmission and spread followed by patient response to vaccines are few of the other topics from this quarter

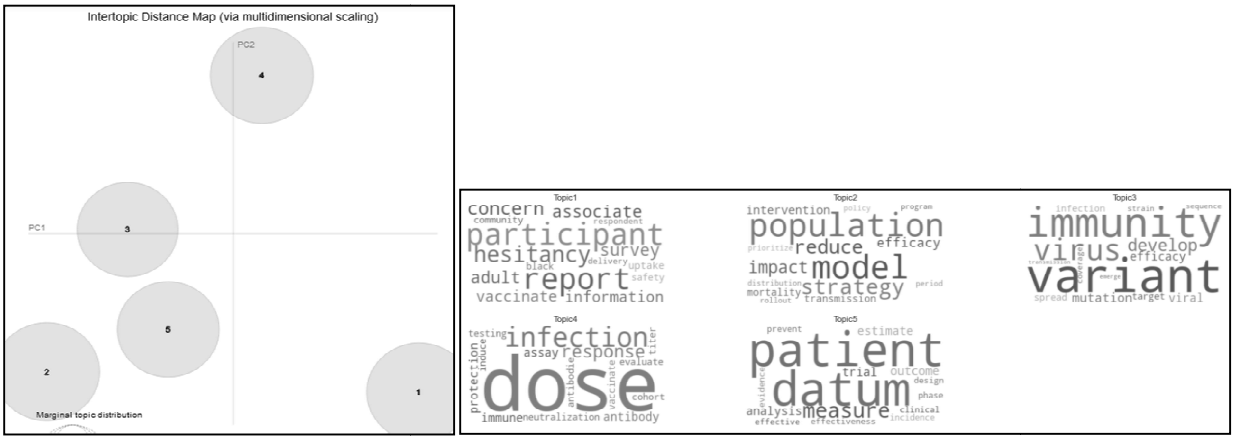


Fig. 6 — Word cloud for the topics generated for preprints from first quarter of year 2021

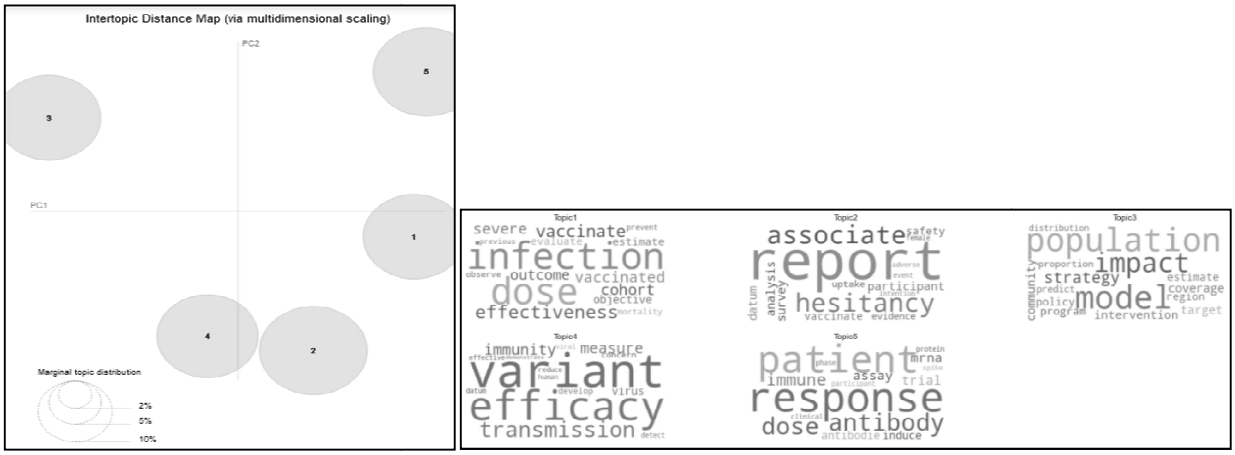


Fig. 7 — Word cloud for the topics generated for preprints from second quarter of year 2021

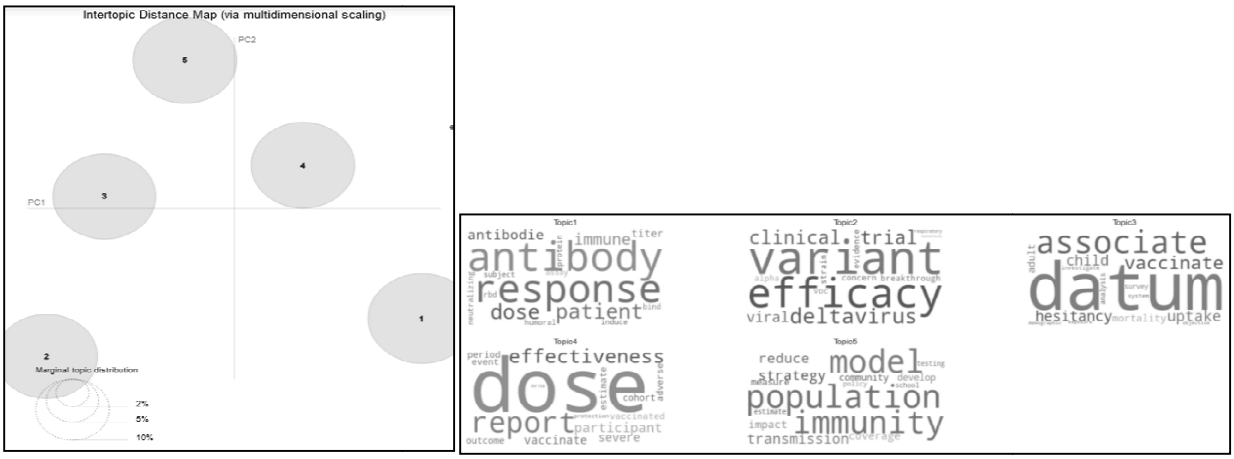


Fig. 8 — Word cloud for the topics generated for preprints from third quarter of year 2021

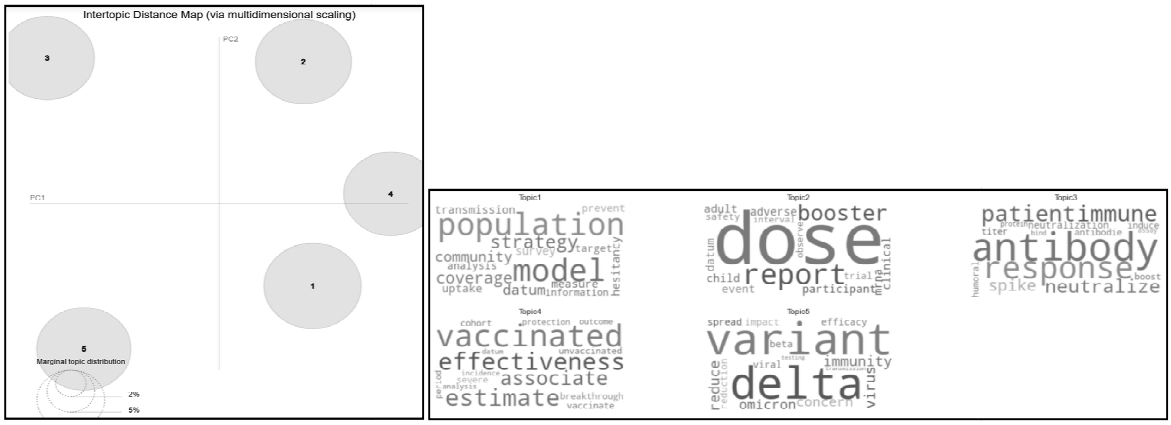


Fig. 9 — Word cloud for the topics generated for preprints from fourth quarter of year 2021

Table 4 – Keywords for each generated topic generated during preprints published in first quarter of year 2021 and the labels assigned to the topics.

| | Terms per Topic | Topic labels | Document Count | Percentage |
|---|---|---|---|---|
| Topic1 | participant, report, hesitancy, concern, associate, survey, adult, information, vaccinate, uptake, community, safety, black, respondent, delivery | Vaccination acceptance / concern | 73 | 22.46 |
| Topic2 | population, model, strategy, reduce, impact, efficacy, intervention, mortality, transmission, distribution, period, program, prioritize, policy, rollout | Models for assessing spread, transmission | 88 | 27.08 |
| Topic3 | variant, immunity, virus, develop, mutation, efficacy, viral, spread, target, infection, sequence, coverage, strain, transmission, emerge | Viral variants and mutations | 41 | 12.62 |
| Topic4 | dose, infection, response, antibody, assay, immune, protection, neutralization, evaluate, antibodie, cohort, vaccinate, testing, titer, induce | Response to vaccines | 73 | 22.46 |
| Topic5 | patient, datum, measure, analysis, estimate, trial, outcome, effectiveness, clinical, incidence, design, phase, prevent, evidence, effective | Clinical trial | 50 | 15.38 |
| The dominant topic during first quarter of year 2021 was observed to be Models for assessing spread transmission, Vaccination acceptance/concern followed by patient response to vaccines. | | | | |

Table 5–Keywords for each generated topic generated during preprints published in second quarter of year 2021 and the labels assigned to the topics

| | Terms per topic | Topic label | Document count | Percentage |
|---|---|---|---|---|
| Topic1 | infection, dose, effectiveness, vaccinate, vaccinated, cohort, severe, outcome, objective, estimate, evaluate, mortality, observe, previous, prevent | Vaccine effectiveness | 59 | 15.09 |
| Topic2 | report, hesitancy, associate, participant, analysis, datum, survey, vaccinate, safety, uptake, evidence, intention, event, adverse, female | Vaccination acceptance / concern | 91 | 23.27 |
| Topic3 | population, model, impact, strategy, coverage, intervention, community, target, policy, estimate, program, region, proportion, predict, distribution | Models for vaccination coverage distribution etc | 66 | 16.88 |
| Topic4 | variant, efficacy, transmission, immunity, measure, virus, develop, concern, effective, datum, viral, detect, reduce, human, demonstrate | Vaccine efficacy against virus variants (variants of concern) | 58 | 14.83 |
| Topic5 | response, patient, antibody, dose, immune, assay, trial, mrna, antibodie, induce, participant, clinical, protein, spike, phase | Clinical trial | 117 | 29.92 |
| The dominant topic during second quarter of year 2021 was observed to be Clinical trials, followed by acceptance of vaccination | | | | |

Table 6–Keywords for each generated topic generated during preprints published in third quarter of year 2021 and the labels assigned to the topics

| | Terms per Topic | Topic labels | Document count | Percentage |
|---|---|---|---|---|
| Topic1 | antibody, response, patient, dose, immune, antibodie, titer, subject, induce, neutralizing, protein, assay, rbd, humoral, bind | Response to vaccines | 59 | 15.09 |
| Topic2 | variant, efficacy, delta, virus, clinical, trial, viral, breakthrough, concern, evidence, alpha, voc, strain, respiratory, mutation | Vaccine efficacy against virus variants (variants of concern) | 91 | 23.27 |
| Topic3 | datum, associate, vaccinate, uptake, hesitancy, child, mortality, adult, survey, analysis, investigate, system, demographic, exposure, objective | Vaccination acceptance / concern | 66 | 16.88 |
| Topic4 | dose, report, effectiveness, participant, vaccinate, severe, vaccinated, outcome, period, cohort, estimate, event, adverse, protection, mrna | Clinical trials | 58 | 14.83 |
| Topic5 | population, model, immunity, transmission, reduce, strategy, coverage, impact, community, develop, measure, estimate, school, testing, policy | Models of vaccine coverage and assessing transmission of infection | 117 | 29.92 |
| The dominant topic during third quarter of year 2021 was observed to be Models related to various aspects of vaccination such as coverage, followed by efficacy against variants of concern and breakthrough infections | | | | |

Table 7–Keywords for each generated topic generated during preprints published in fourth quarter of year 2021 and the labels assigned to the topics

| | Terms per Topic | Topic labels | Document Count | Percentage |
|---|---|---|---|---|
| Topic1 | population, model, strategy, coverage, community, datum, transmission, measure, hesitancy, uptake, survey, prevent, analysis, target, information | Models for vaccine coverage, infection spread etc | 121 | 23.27 |
| Topic2 | dose, report, booster, participant, adult, mrna, child, event, clinical, adverse, datum, safety, trial, interval, observe | Clinical trials | 73 | 14.04 |
| Topic3 | antibody, response, patient, immune, neutralize, spike, neutralization, antibodie, titer, induce, boost, humoral, protein, bind, assay | Patient response | 140 | 26.92 |
| Topic4 | vaccinated, effectiveness, estimate, associate, breakthrough, severe, vaccinate, cohort, unvaccinated, protection, outcome, period, analysis, incidence, datum | Estimating vaccination effectiveness | 99 | 19.04 |
| Topic5 | variant, delta, immunity, reduce, virus, omicron, concern, efficacy, impact, viral, spread, reduction, beta, transmission, testing | Efficacy of vaccines in light of variants of concern such as delta and Omicron | 87 | 16.73 |
| The dominant topic during last quarter of year 2021 was observed to be patient responses to vaccines, followed by models for transmission and vaccine coverage. This quarter is also marked by publications of articles related to vaccine efficacy considering variants of concern such as Delta and Omicron | | | | |

to induction of immunity or boosting of immune responses after vaccine administration. The third important cluster was observed to be of articles that communicated findings from various phases of the clinical trials involving candidate vaccines against COVID-19. The articles related to concerns or hesitancy of the public regarding vaccination also form one of the important clusters. The keywords generated in 5 topics of each of the 7 quarters were analysed for occurrence of the same keywords in the 5 topics of each quarter. It was observed that at least 31 keywords were occurring in different topics of different quarters and such occurrence was found in at least 5 combinations of 7 quarters and 5 topics. Furthermore, it was also observed that keywords like ‘analysis’ and ‘estimate’ occurred 9 times, ‘transmission’ and ‘vaccinate’ occurred 8 times while 'report', 'associate', 'participant', 'survey', 'clinical', 'virus', 'community', 'immune', 'viral', 'antibodie', 'datum' and 'efficacy' were present at least 7 times in the combinations of 7 quarters and 5 topics.

A careful analysis of the topics in each quarter reflected that few topics are present consistently across quarters, indicating their importance. Such consistent topics include clinical trials and patient responses to the vaccines. Vaccine hesitance also has been a consistently occurring theme since the last quarter of year 2020 and combating vaccine hesitancy is important to overcome the pandemic with help of vaccination. As there are topics that are consistent across quarters, we also observed that some topics emerged later such as vaccine efficacy in light of viral mutations and emergence of variants of concern, which have been more prominent from the first quarter of the year 2021, with the Omicron variant also dominating this cluster in the last quarter.

## Limitations of the study

There are multiple preprint repositories available, and the present work highlights the contents from one of the several preprint repositories. While many have appreciated the quality of work deposited in preprint servers, few have also debated the authenticity and quality of work as these publications are raw and not peer reviewed ([2,5,6,7,19]). Hence a cautious approach is recommended while dealing with the content on preprint repositories. The topic modelling carried out is based on the predefined algorithms with little or no control on it. It was observed that the NLP SpaCy python library which was used for tokenization and lemmatization of the unique words did not work effectively and yielded two similar keywords 'antibody', and 'antibodie' in one of the topics for all the 7 quarters under analysis.

## Conclusion

MedRxiv preprint played an important role in communicating the research results on vaccines in response to the COVID-19 pandemic. The researchers not only communicated the results from clinical trials of vaccines but articles on other relevant topics such as vaccination models and vaccine hesitancy have also been published in the MedRxiv platform and the preprints have been agile in responding to the pandemic and it is felt that the preprints are here to stay even if pandemic doesn't.

## References


1 Canouï, E., and Launay O. , "Histoire et principes de la vaccination." *Revue des maladies respiratoires* 36(1) (2019): 74-81. DOI: 10.1016/j.rmr.2018.02.015

2 Chung, K. J., "Preprints: What is their role in medical journals?." *Archives of Plastic Surgery* 47(0)2 (2020): 115-117. DOI: 10.5999/aps.2020.00262

3 Fry, N K., Helina M., and Tasha M. C., "In praise of preprints." *Access Microbiology* 1(2) (2019). DOI: 10.1099/acmi.0.000013

4 Älgå, A., Oskar E., and Nordberg M., "The development of preprints during the COVID-19 pandemic." *Journal of Internal Medicine* 290(2) (2021): 480. DOI: 10.1111/joim.13240

5 Bloom, T. "Shepherding preprints through a pandemic." *BMJ: British Medical Journal* 371 (2020): m4703. DOI: 10.1136/bmj.m4703

6 van Schalkwyk, M. C., Hird, T. R., Maani, N., Petticrew, M., and Gilmore, A. B. (2020). "The perils of preprints." *BMJ: British Medical Journal (Online)* 370 (2020). DOI: 10.1002/alr.22732

7 Hopkins, Claire. "Preprints—expediting access or compromising quality?." *International Forum of Allergy & Rhinology*. 11(5). 2021.

8 Blei, D. M., Andrew Y. N., and Jordan M. I. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.

9 Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T. and Schmid-Petri, H., "Applying LDA topic modelling in communication research: Toward a valid and reliable methodology." *Communication Methods and Measures* 12(2-3) (2018): 93-118. DOI: 10.1080/19312458.2018.1430754

10 Elgesem, D., Steskal, L. and Diakopoulos, N., "Structure and content of the discourse on climate change in the blogosphere: The big picture." *Environmental Communication* 9(2) (2015): 169-188. DOI: 10.1080/17524032.2014.983536

11 Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. "Public sentiment analysis and topic modelling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence." *Journal of Infection and Public Health* 14(10) (2021): 1505-1512. DOI: 10.1016/j.jiph.2021.08.010

12 Dong, M., Cao, X., Liang, M., Li, L., Liu, G., & Liang, H. "Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modelling." *MedRxiv* (2020): 2020-03. DOI: 10.1101/2020.03.26.20044164

13 Johansson, R., & Engström Heino, O. Topic propagation over time in internet security conferences: Topic modelling as a tool to investigate trends for future research. Bachelors Dissertation. Linköping University, Sweden, 2021.

14 McCallum, A. K., (2002), "MALLET: A Machine Learning for Language Toolkit." Available at https://mimno.github.io/Mallet/about (Accessed on 6 March 2023)

15 He J, Larson M, and De Rijke M."Using coherence-based measures to predict query difficulty." *Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings 30*. Springer Berlin Heidelberg, 2008.

16 Dong, M., Cao, X., Liang, M., Li, L., Liu, G. and Liang, H., "Understand research hotspots surrounding COVID-19 and other coronavirus infections using topic modelling." *MedRxiv* (2020): 2020-03. DOI: 10.1101/2020.03.26.20044164

17 Stevens, K., Kegelmeyer, P., Andrzejewski, D. and Buttler, D., "Exploring topic coherence over many models and many topics." *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. 2012.

18 Abhishek, T., Rawat, D., Gupta, M. and Varma, V., "Transformer models for text coherence assessment." *arXiv preprint arXiv:2109.02176* (2021). DOI: 10.48550/ arXiv.2109.02176

19 Fraser, N., Brierley, L., Dey, G., Polka, J.K., Pálfy, M., Nanni, F. and Coates, J.A., "The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape." *PLoS Biology* 19(4) (2021): e3000959. DOI: 10.1371/journal.pbio.3000959