

ORIGINAL ARTICLE

Latent Dirichlet allocation topic modeling of free-text responses exploring the negative impact of the early COVID-19 pandemic on research in nursing

Madoka Inoue^{1,2}  | Hiroki Fukahori^{3,4} | Manami Matsubara⁵ |
Naoki Yoshinaga^{1,4}  | Hideo Tohira^{1,2,6}

¹University of Miyazaki, Miyazaki, Japan

²Curtin University, Bentley, Western Australia, Australia

³Keio University, Tokyo, Japan

⁴COVID-19 Nursing Research Countermeasures Committee, Japan Academy of Nursing Science, Tokyo, Japan

⁵Kansai International University, Miki, Japan

⁶The University of Western Australia, Crawley, Western Australia, Australia

Correspondence

Madoka Inoue, Curtin Medical School
Kent Street, Curtin University, Bentley,
WA, 6102, Australia.

Email: madoka.inoue@curtin.edu.au

Abstract

Aim: To derive latent topics from free-text responses on the negative impact of the pandemic on research activities and determine similarities and differences in the resulting themes between academic-based and clinical-based researchers.

Methods: We performed a secondary analysis of free-text responses from a cross-sectional online survey conducted by the Japan Academy of Nursing Science of its members in early 2020. The participants were categorized into two groups by workplace (academic-based and clinical-based researchers). Latent Dirichlet allocation (LDA) topic modeling was used to extract latent topics statistically and list important keywords/text associated with the topics. After organizing similar topics by principal component analysis (PCA), we finally derived topic-associated themes by reading the keywords/texts and determining the similarity and differences of the themes between the two groups.

Results: A total of 201 respondents (163 academic-based and 38 clinical-based researchers) provided free-text responses. LDA identified eight and three latent topics for the academic-based and clinical-based researchers, respectively. While PCA re-grouped the eight topics derived from the former group into four themes, no merging of the topics from the latter group was performed resulting in three themes. The only theme common to the two groups was “barriers to conducting research,” with the remaining themes differing between the groups.

Conclusions: Using LDA topic modeling with PCA, we identified similarities and differences in the themes described in free-text responses about the negative impact of the pandemic between academic-based and clinical-based researchers. Measures to mitigate the negative impact of pandemics on nursing research may need to be tailored separately.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Japan Journal of Nursing Science* published by John Wiley & Sons Australia, Ltd on behalf of Japan Academy of Nursing Science.

KEYWORDS

latent Dirichlet allocation, topic modeling, principal component analysis, nursing research, negative impact

1 | INTRODUCTION

With the high infectivity and morbidity of the novel coronavirus disease (COVID-19), the first few months of the pandemic saw restrictions imposed on clinical sites at the request of the Japanese government, thereby making clinical research challenging (Prime Minister of Japan and His Cabinet, 2020). Many nursing research activities were blocked or delayed, especially in terms of data collection, as it often requires direct interaction with patients. In light of these circumstances, the Japan Academy of Nursing Science (JANS) cross-sectionally surveyed its members online in early 2020 about the negative impacts of the pandemic on their research activities to ascertain areas in need of support (Japan Academy of Nursing Science, 2020). Since the JANS members mainly comprise researchers engaging in either academic or clinical activities, and each research environment is different, it is essential to understand their needs separately. Nevertheless, only one study investigated the negative impacts of the pandemic on research activities by comparing the above two groups in Japan. Inoue et al. (2022) determined that consulting support on information and communication technology (ICT)-related issues for academic-based researchers, as well as counseling for research concerns for clinical-based researchers, were urgently needed. However, that study analyzed only pre-defined responses. Hitherto, few studies have explored free-text responses to identify the negative impacts on research that pre-defined responses may not capture.

There are significant discrepancies between free-text and pre-defined responses to the same question (Ogden & Lo, 2012). Specifically, free-text responses usually provide more detailed and broader information than pre-defined ones (Friborg & Rosenvinge, 2011). Despite this, the analysis of free-text responses is often costly and laborious as it involves several steps by multiple coders, including the development of a categorization scheme, coder training, coding, and a reliability check (Züll, 2016). Due to these complexities, the analysis of free-text responses is prone to inefficiency in coding and disagreement among coders.

Today, as digitized data are more available and text-mining software becomes more accessible than before, many researchers have come to use topic modeling (Buenano-Fernandez et al., 2020; Chung et al., 2022; Pietsch & Lessmann, 2019; Vijayan, 2021). Topic modeling is one of the natural language processing (NLP) techniques in the field of machine learning. It considers that

documents include a mixture of latent topics and provides each document with probabilities that the document is associated with each topic using various statistical methods such as latent Dirichlet allocation (LDA). LDA has been used for a variety of research in topic extraction (Bashri & Kusumaningrum, 2017; Onan, 2019; Onan et al., 2016). Although there are other methods in topic modeling, including Latent Semantic Index and Probabilistic Latent Semantic Indexing, LDA provides better topic coherence than others (Garbhapu & Bodapati, 2020). Since topic modeling can be performed by computer software, this technique may make up for the shortcomings of traditional methods that analyze free-text responses. Nonetheless, few nursing studies have applied this technique, especially to fully disclose the impact on research activities posed by the pandemic.

This study aimed to derive latent topics from free-text responses to the JANS survey question about the negative impact of the early COVID-19 pandemic on nursing research, and determine the similarities and differences between academic-based and clinical-based researchers in the themes of topics in their responses.

2 | METHODS

2.1 | Study setting and period

We performed a secondary analysis of free-text responses from a cross-sectional online survey conducted by the JANS of its members (Japan Academy of Nursing Science, 2020). The JANS survey was written in Japanese and investigated the impacts on research activities experienced during the first few months of the COVID-19 pandemic, and was distributed between July 1 and August 10, 2020. The details of this JANS online survey with the statement of the Checklist for Reporting Results of Internet E-Surveys (CHERRIES) were published elsewhere by the authors (Inoue et al., 2022).

2.2 | Inclusion criteria and subjects of this study

Among all participants, we included those who provided free-text responses to the following survey question: “Factors that have negatively impacted your research activities during the COVID-19 pandemic.” Based on their workplaces, we first classified the

included participants into two groups (academic-based and clinical-based researchers). For example, the former group includes those working in academic environments (e.g., university, college and/or research institute), and the latter was those working in clinical environments (e.g., hospitals, clinics and/or care facilities). We classified those who did not provide a type of workplace in the latter group.

2.3 | The investigative process of free-text responses

Figure 1 shows a flowchart that outlines the processing of free-text responses to the question, which takes 10 steps, including LDA topic modeling and principal component analysis (PCA). The details of LDA and PCA are described in a later section. In this study, we began by extracting free-text responses to the question and processed the responses for each participant group separately (Figure 1, Step 1).

2.4 | Data pre-processing steps

We pre-processed the extracted free-text responses using Python and its libraries as follows. In contrast to English or other Western languages, words are not separated by spaces in Japanese. Therefore, texts need to be segmented first into words before further analysis (Figure 1, Step 2). We used Janome (version 0.4.1) for this purpose (“Janome v0.4 documentation (en)”, 2020). After the text segmentation, nouns were extracted (Figure 1, Step 3), and all punctuation and pre-selected stop words, a set of commonly used words with little meaningful information (e.g., 私 [the first person singular pronoun in English]), were removed (Figure 1, Step 4). After this removal, we derived a corpus including words (unigram) and two and three consecutive words (bigrams and trigrams, respectively) using the remaining words (Figure 1, Step 5). We collectively refer to unigram, bigram, and trigram as ngrams hereafter. Finally, we computed the term frequency-inverse document frequency (tf-idf) weights to quantify the importance of ngrams (Jurafsky & Martin, 2009) (Figure 1, Step 6). The tf-idf is the most commonly used weighting method of ngrams in NLP (Beel et al., 2015). The value of a tf-idf weight is a product of the frequency of an ngram in a text and a logarithmized inverse of the proportion of texts in the corpus that contain the ngram. Hence, the value of the tf-idf weight is greater when an ngram frequently appears in a text but does not appear in other texts. We separately derived sets of tf-idf weights for academic-based and clinical-based researchers using Python's gensim library (version 4.1.2).

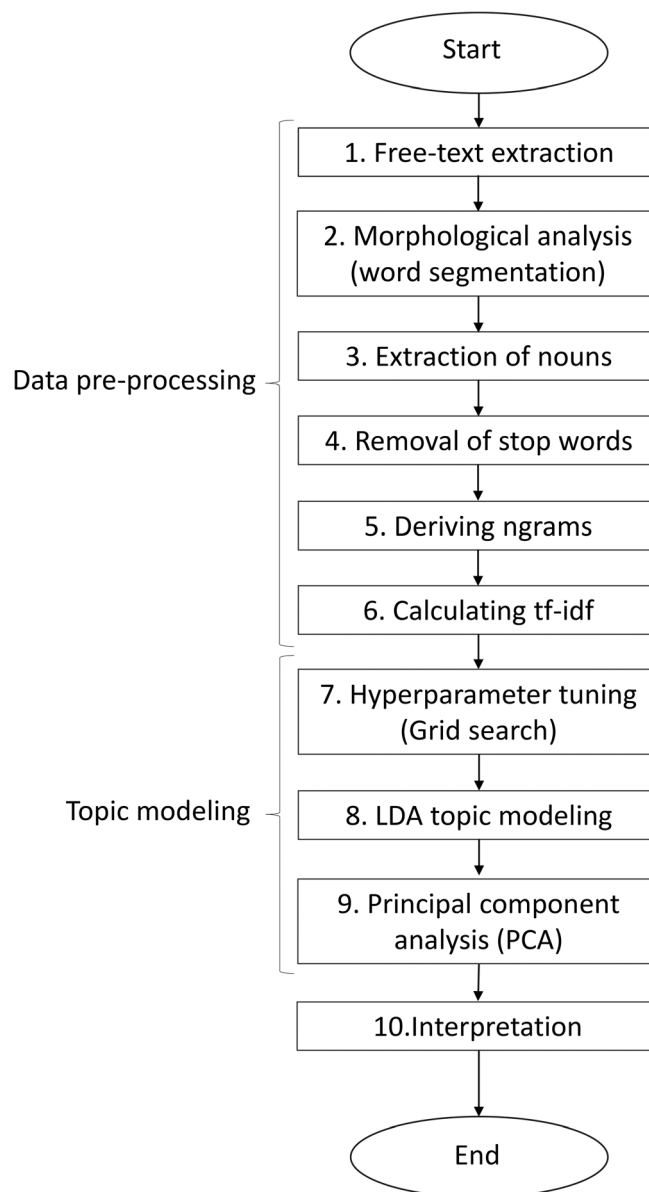


FIGURE 1 Flowchart of data processing and modeling. LDA, latent Dirichlet allocation; tf-idf, term frequency-inverse document frequency

2.5 | Latent Dirichlet allocation (LDA) topic modeling and principal component analysis (PCA)

We performed topic modeling with the tf-idf weights. Topic modeling is the mathematical method for NLP, including classifying words with similar patterns to infer latent documented topics. We used the LDA for this purpose (Blei et al., 2003). LDA is known as an unsupervised method and considers that each ngram is associated with a specific latent topic and uses the ngrams to determine the topics of documents. For example, “tire,” “handle,” and “wiper” are likely to be associated with cars;

therefore, documents in which these words frequently appear can be considered to include a topic about cars. If multiple topics were identified in a particular response, LDA provides probabilities associated with the topics. LDA also provides each ngram with probabilities that the ngram belongs to various topics.

Three parameters need to be set to gain optimal topic modeling by LDA. These parameters include the number of latent topics included in the text and two hyperparameters (alpha and beta) that determine the shape of Dirichlet distribution. Because we neither know how many latent topics are included nor what values for the hyperparameters are the most appropriate, we derived as many as 540 LDA models by changing the number of topics and values of the two hyperparameters for each participant group (Figure 1, Step 7). We used coherence as a performance measure to determine the best-performing LDA models. Coherence represents the similarity between words in a given topic (Mimno et al., 2011). The greater the coherence is, the better the model becomes. We selected a model with hyperparameters by which an LDA model yielded the greatest coherence. This process was repeated for each group. We integrated similar topics yielded from the LDA models by PCA (Figure 1, Step 9). PCA is a dimensionality reduction technique to extract core components with as little loss of the original information as possible (Salih Hasan & Abdulazeez, 2021). We applied PCA to ease the data visualization and avoid the curse of dimensionality, which may produce unreliable results (Verleysen & Francois, 2005). An intertopic distance map was drawn to demonstrate the distance among the topics on a 2-dimensional plane using pyLDAvis (version 3.3.1) (Figure 3). On the map, each topic was drawn in the space by a circle. The radius of the circles is proportional to the word amount belonging to the topic in the free-text responses. The distance between circles shows the closeness between the topics. The closer the circles are, the more similar are the topics. We considered that topics were similar if circles overlapped and integrated topics of those circles into a single group. This process usually is only performed if the sample size of the groups is 40 or greater (Shaukat et al., 2016). In this study, we undertook this process only for academic-based researchers due to their sample including over 40 members.

2.6 | Interpretation

After the integration of the topics, we determined the themes of topics by reading actual texts and keywords associated with the topics (Figure 1, step 10). All authors

independently conducted this process. Disagreements among the authors were solved by consensus. Keywords belonging to a given topic (topic keywords) were drawn in bar charts in descending order of topic-word probability, the probability that a keyword is likely to belong to that topic (Figure 2). We also summarized each topic using the keywords in Figure 2 corresponding to the actual texts. Word clouds were generated to present the keywords of a topic visually. In the word clouds, the size of the word is proportional to the value of the tf-idf of the word in the texts. Hence, words drawn in a large size in a word cloud for a given topic are exclusively used in texts belonging to the topic but less common in other topics. Due to the nature of the original language used in the survey, that is, Japanese, the results of word cloud generation are shown in Appendix S1 in the Supporting Information. Similarities and differences in the themes between the participant groups were determined based on the groups of words without being based on any quantity after clustering groups of words using the vectors.

2.7 | Analysis

We performed Chi-square or Fisher's exact tests, where appropriate, to compare baseline characteristics of the respondents, and set the significance level at .05. We excluded cases with missing data in the comparisons. We used Python (version 3.8.11) to manage data and the gensim library (version 3.8.3) for NLP, including pre-processing and LDA modeling.

2.8 | Ethical consideration

This study was approved by the Research Ethics Committee of the University of Miyazaki (Approved Number: O-0733). Due to the nature of the study design, we also held multiple discussions with the COVID-19 Nursing Research Countermeasures Committee of JANS, and with the other five research groups that planned to use the JANS data for their studies to avoid duplicated or data-fragmented publications. Currently, four articles have been published (https://www.jans.or.jp/modules/en/index.php?content_id=80#covid-19pu). We confirmed that each study had different research questions, methods, subjects, and analyses, and finally obtained permission from the JANS committee to carry out the analysis of this study (https://www.jans.or.jp/modules/en/index.php?content_id=80#covid-19pu) (Japan Academy of Nursing Science, 2020).

(a) The 10 most associated words with each topic for academic-based researchers

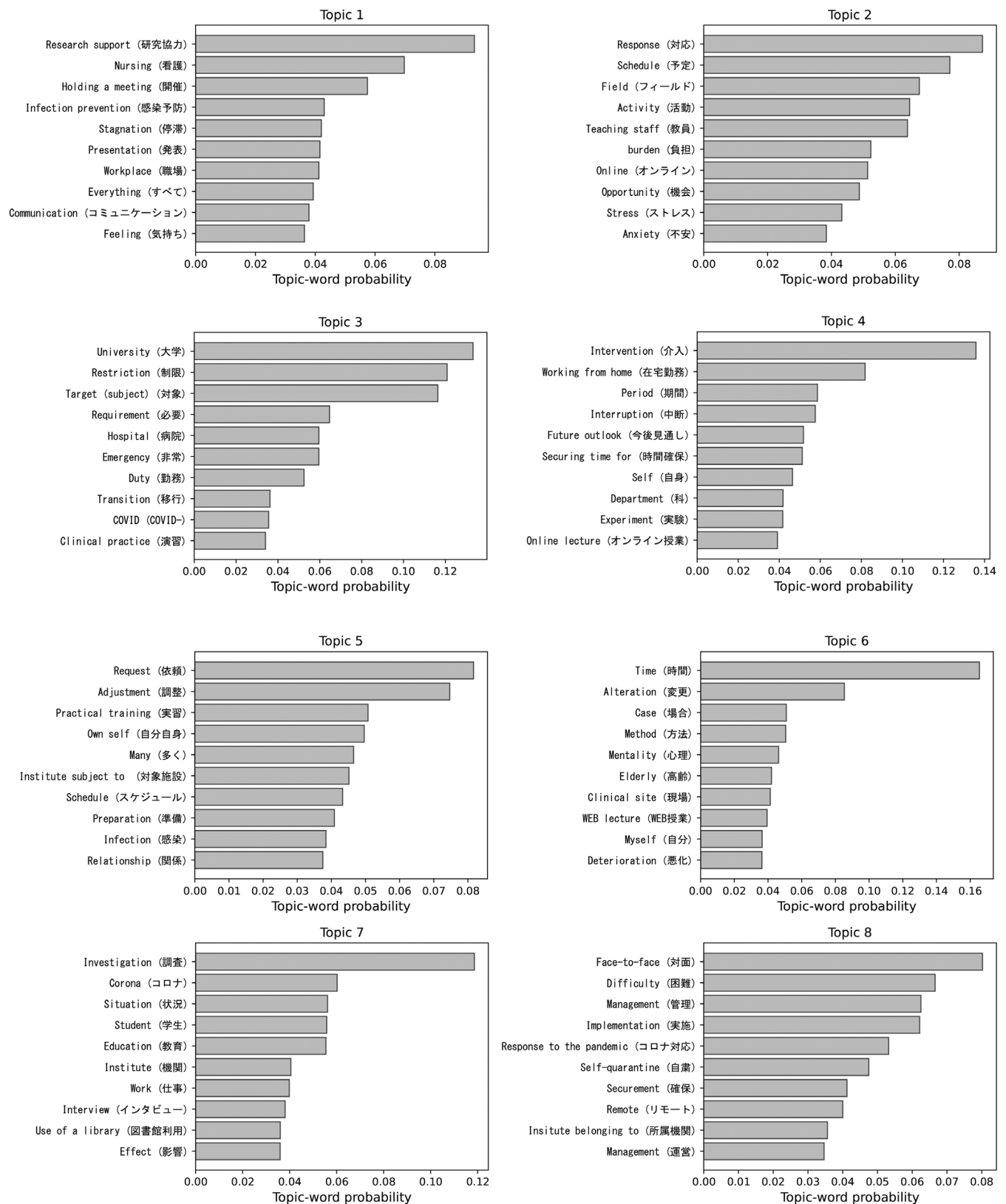


FIGURE 2 (a) The 10 most associated words with each topic for academic-based researchers. (b) The 10 most associated words with each topic for clinical-based researchers

(b) The 10 most associated words with each topic for clinical-based researchers

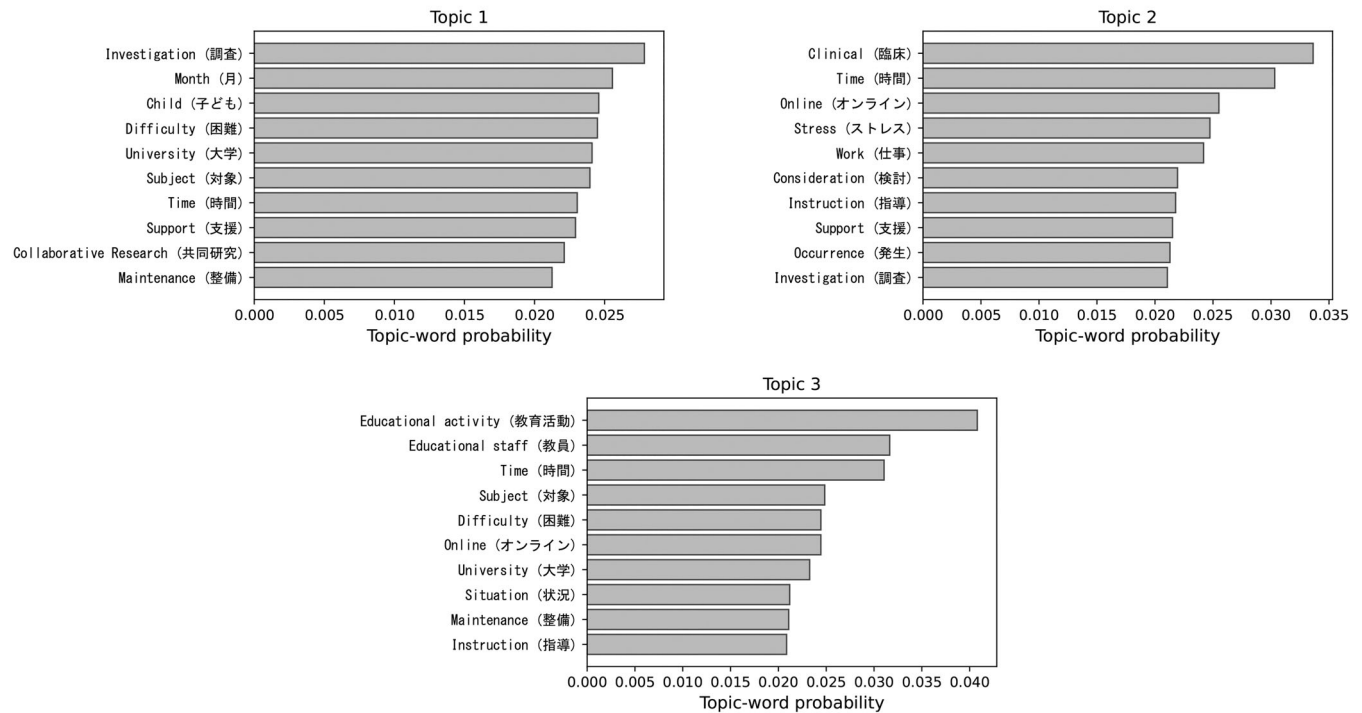


FIGURE 2 (Continued)

3 | RESULTS

3.1 | Baseline characteristics

Among 9524 JANS members, 1532 completed the survey with a response rate of 16.8%. Of the 1532 respondents, 201 (163 academic-based and 38 clinical-based researchers) provided free-text responses to the above question (Table 1). Those aged 46–55 years accounted for 37.4%, followed by those aged ≥ 56 years (28.3%) and those aged 36–45 years (25.7%) (Table 1). Females were dominant (90.8%), with 72.1% of the respondents living in one of the prefectures under the special precautions against COVID-19. Most of the baseline characteristics were similar between the two groups, except for the types of workplace ($p < .001$), job titles ($p < .001$), and the distribution of the attained highest academic degree ($p = .01$) (Table 1).

3.2 | Topic modeling

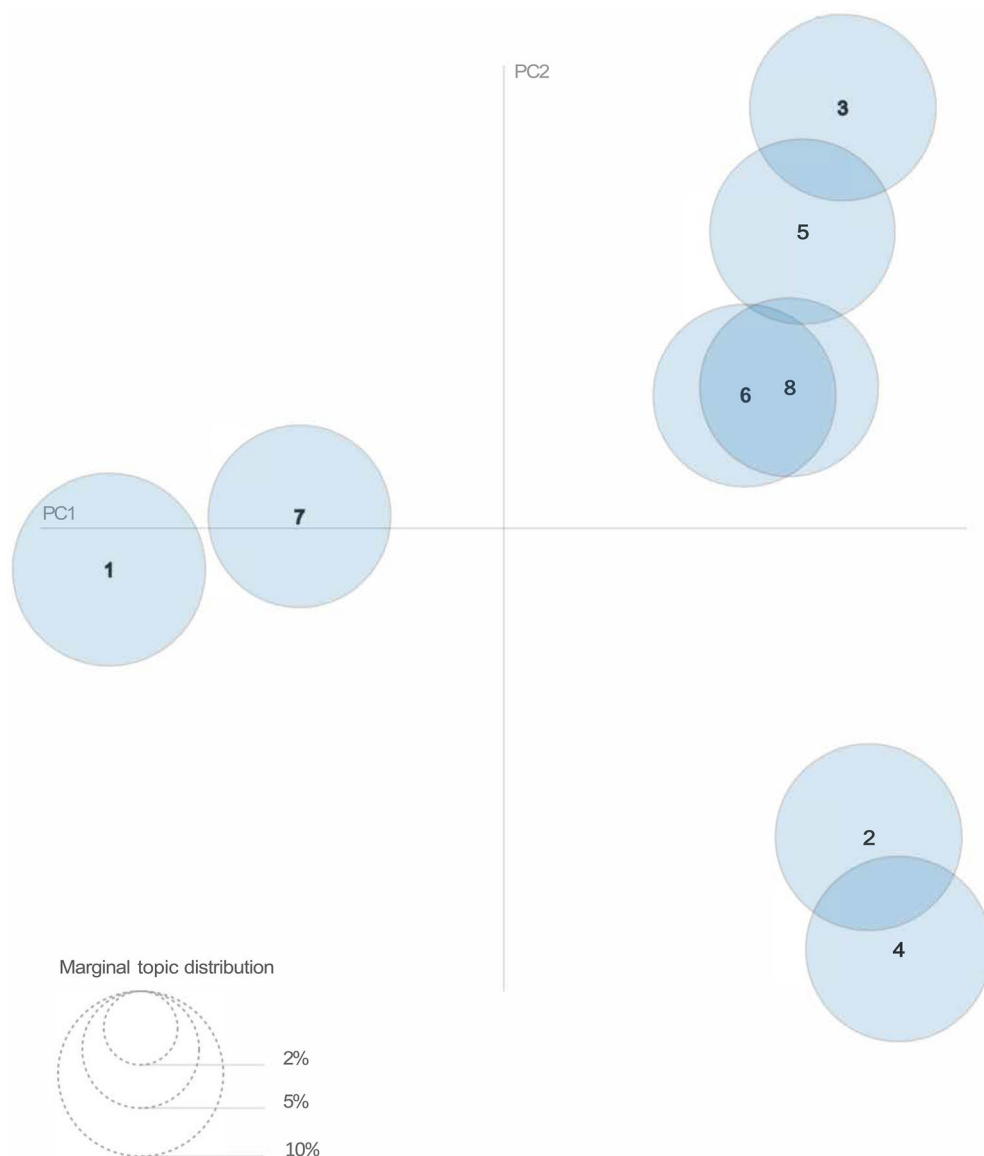
Figure 2 presents the word probabilities of the top 10 words in each latent topic of the two participant groups. The LDA found eight latent topics in the free-text responses for academic-based researchers (Figure 2a) and three topics for clinical-based researchers (Figure 2b). While the top five examples of descriptions in each topic by workplace written in Japanese are provided in

Appendix S2, English summaries of each topic are as follows.

3.2.1 | Academic-based researchers

- **Topic 1** described that research-related administrative areas, including the ethics committee and research governance department, stopped their activities, slowing the progression of research. It also described that many meetings/conferences were canceled. The participants felt that everything had stagnated. Preventative measures against the pandemic implemented at workplaces led to losing communication with collaborators and other people.
- **Topic 2** was related to the extra burden placed on participants' work that evoked negative emotions. The participants were anxious and exhausted from responding to the changes in providing classes online and in their schedules. Some were frustrated because they could not go to their study sites to conduct research. These changes or responses created an extra burden.
- **Topic 3** was mainly about restrictions implemented at the workplace, including in universities and hospitals. In particular, the participants who resided in the areas under the emergency alert could not go to these facilities or required special permission when entering clinical placement sites or contacting their study subjects and targets.

FIGURE 3 Results of principal component analysis for academic-based researchers. Numbers represent the topics identified in responses by latent Dirichlet allocation



- **Topic 4** mentioned the uncertainty of future outlook. The participants became capable at working from home and giving lectures online. However, their own studies, experiments, or library use were limited or interrupted, provoking apprehension and confusion about their futures.
- **Topic 5** was about the many changes in their own lives. Some stated that they needed to adjust their schedules of clinical practice to respond to the requests from clinical sites because of the efforts at infection control.
- **Topic 6** included keywords relating to various measures against the pandemic that had a deleterious impact on research activities, resulting in work overload. These include changes involving time, changes in the method of teaching from face-to-face to web-based lectures, and the cases they were allocated at clinical sites.
- **Topic 7** focused on the disruptions of research activities by the pandemic. Some described that the recruitment

of study subjects for interviews was suspended or the library was inaccessible; others commented on the increased educational workload for students.

- **Topic 8** was associated with the impacts and difficulties of implementing preventive measures and keeping social and physical distance. The participants had to deal with the protocols of coronavirus measures and follow the rules formulated at workplaces to prevent the spread of coronavirus infection.

3.2.2 | Clinical-based researchers

- **Topic 1** described obstacles to research activities. Some stated that they had to stay home with their children due to temporary closures and thus could not go to a university or clinical site.
- **Topic 2** was related to increased clinical duties. The participants had to deal with the infection

TABLE 1 Baseline characteristics of the respondents

		Academic-based researchers	Clinical-based researchers	All	p-value
Total	N	163	38	201	
Age	≤35	12, 7.5% (3.4%–11.5%)	4, 15.4% (1.5%–29.3%)	16, 8.6% (4.5%–12.6%)	.158
	36–45	43, 26.7% (19.9%–33.5%)	5, 19.2% (4.1%–34.4%)	48, 25.7% (19.4%–31.9%)	
	46–55	57, 35.4% (28.0%–42.8%)	13, 50.0% (30.8%–69.2%)	70, 37.4% (30.5%–44.4%)	
	≥56	49, 30.4% (23.3%–37.5%)	4, 15.4% (1.5%–29.3%)	53, 28.3% (21.9%–34.8%)	
	Unknown	2	12	14	
Sex	Female	145, 91.2% (86.8%–95.6%)	23, 88.5% (76.2%–100.0%)	168, 90.8% (86.6%–95.0%)	.65
	Male	14, 8.8% (4.4%–13.2%)	3, 11.5% (0.0%–23.8%)	17, 9.2% (5.0%–13.4%)	
	Unknown	4	12	16	
Workplace	University	161, 98.8% (97.1%–100.0%)	0	161, 85.6% (80.6%–90.7%)	<.001
	Research institute	2, 1.2% (0.0%–2.9%)	0	2, 1.1% (0.0%–2.5%)	
	Hospital/clinic	0	17, 68.0% (49.7%–86.3%)	17, 9.0% (4.9%–13.1%)	
	Others	0	4, 16.0% (1.6%–30.4%)	4, 2.1% (0.1%–4.2%)	
	Unemployed	0	4, 16.0% (1.6%–30.4%)	4, 2.1% (0.1%–4.2%)	
	Unknown	0	13	13	
Job title	Professor	46, 28.4% (21.5%–35.3%)	1, 4.0% (0.0%–11.7%)	47, 25.1% (18.9%–31.4%)	<.001
	Associate professor	33, 20.4% (14.2%–26.6%)	0	33, 17.6% (12.2%–23.1%)	
	Lecturer	40, 24.7% (18.1%–31.3%)	1, 4.0% (0.0%–11.7%)	41, 21.9% (16.0%–27.9%)	
	Assistant professor	37, 22.8% (16.4%–29.3%)	0	37, 19.8% (14.1%–25.5%)	
	Teaching associate	3, 1.9% (0.0%–3.9%)	1, 4.0% (0.0%–11.7%)	4, 2.1% (0.1%–4.2%)	
	Nurse manager	0	8, 32.0% (13.7%–50.3%)	8, 4.3% (1.4%–7.2%)	
	Full-time clinical nurse	0	3, 12.0% (0.0%–24.7%)	3, 1.6% (0.0%–3.4%)	
	Part-time clinical nurse	0	3, 12.0% (0.0%–24.7%)	3, 1.6% (0.0%–3.4%)	
	Other	0	0	0	
	Unknown	1	13	14	
Employment type	Full-time (fixed term)	65, 40.4% (32.8%–48.0%)	5, 20.0% (4.3%–35.7%)	70, 37.6% (30.7%–44.6%)	<.001
	Full-time (permanent)	93, 57.8% (50.1%–65.4%)	13, 52.0% (32.4%–71.6%)	106, 57.0% (49.9%–64.1%)	
	Part-time	3, 1.9% (0.0%–4.0%)	5, 20.0% (4.3%–35.7%)	8, 4.3% (1.4%–7.2%)	
	Other	0	2, 8.0% (0.0%–18.6%)	2, 1.1% (0.0%–2.6%)	
	Unknown	2	13	15	
Lived in a special alert area	Yes	117, 71.8% (64.9%–78.7%)	20, 74.1% (57.5%–90.6%)	137, 72.1% (65.7%–78.5%)	.81
	No	46, 28.2% (21.3%–35.1%)	7, 25.9% (9.4%–42.5%)	53, 27.9% (21.5%–34.3%)	
	Unknown	0	11	11	

TABLE 1 (Continued)

		Academic-based researchers	Clinical-based researchers	All	p-value
Job change	Yes	16, 9.8% (5.2%–14.4%)	5, 18.5% (3.9%–33.2%)	21, 11.1% (6.6%–15.5%)	.18
	No	147, 90.2% (85.6%–94.8%)	22, 81.5% (66.8%–96.1%)	169, 88.9% (84.5%–93.4%)	
	Unknown	0	11	11	
Highest degree	PhD	84, 51.9% (44.2%–59.5%)	9, 32.1% (14.8%–49.4%)	93, 48.9% (41.8%–56.1%)	.01
	Master	77, 47.5% (39.8%–55.2%)	16, 57.1% (38.8%–75.5%)	93, 48.9% (41.8%–56.1%)	
	Bachelor	1, 0.6% (0.0%–1.8%)	1, 3.6% (0.0%–10.4%)	2, 1.1% (0.0%–2.5%)	
	Other	0	2, 7.1% (0.0%–16.7%)	2, 1.1% (0.0%–2.5%)	
	Unknown	1	10	11	
Early, mid-career researcher	Yes	38, 46.3% (35.5%–57.1%)	5, 55.6% (23.1%–88.0%)	43, 47.3% (37.0%–57.5%)	.58
	No	44, 53.7% (42.9%–64.5%)	4, 44.4% (12.0%–76.9%)	48, 52.7% (42.5%–63.0%)	
	Unknown	2	0	2	
Live with a spouse/partner	Yes	97, 62.2% (54.6%–69.8%)	15, 60.0% (40.8%–79.2%)	112, 61.9% (54.8%–69.0%)	.83
	No	59, 37.8% (30.2%–45.4%)	10, 40.0% (20.8%–59.2%)	69, 38.1% (31.0%–45.2%)	
	Unknown	7	13	20	
Currently raising a child/children	Yes	56, 35.9% (28.4%–43.4%)	8, 32.0% (13.7%–50.3%)	64, 35.4% (28.4%–42.3%)	.71
	No	100, 64.1% (56.6%–71.6%)	17, 68.0% (49.7%–86.3%)	117, 64.6% (57.7%–71.6%)	
	Unknown	7	13	20	
Currently caring for the elderly	Yes	23, 14.7% (9.2%–20.3%)	4, 16.0% (1.6%–30.4%)	27, 14.9% (9.7%–20.1%)	.87
	No	133, 85.3% (79.7%–90.8%)	21, 84.0% (69.6%–98.4%)	154, 85.1% (79.9%–90.3%)	
	Unknown	7	13	20	

countermeasures and protocols of COVID-19. These non-research activities or online work caused stress.

- **Topic 3** focused on the difficulty for teaching activities or learning opportunities because of the limited time and indirect communication. Some described that they consumed considerable portions of their time coordinating with various areas and explaining situations to students and other stakeholders.

The word frequency in the actual text using word clouds is visually represented in Appendix S1. Some words in Figure 2 appeared more prominent than other words because they were mentioned more within the texts.

The PCA was conducted only for academic-based researchers and displayed using the intertopic distance map (Figure 3). Each numbered circle represented a topic from Topic 1 to Topic 8. We found similarities within one group of four topics (Topics 3, 5, 6, and 8) and another of

two topics (Topics 2 and 4), as the circles of the topics in each group overlapped (Figure 3). These two groups of overlapping topics were thus each integrated into a separate theme, resulting in four separate themes from the eight latent topics for academic-based researchers. On the other hand, PCA was not performed for clinical-based researchers because there were fewer than 40 respondents, the final result remained as three themes from three latent topics (Table 2).

Table 2 outlines the themes derived after reading and analyzing the actual free-text responses from academic-based and clinical-based researchers. It also describes the relationships among the topics, topic groups, and themes (Table 2). The theme of “Barriers to conducting research” was commonly derived for both participant groups, while the rest of the themes differed between the groups. Specifically, the academic-based researchers generated three unique themes, comprising “Stagnation of research support environment,” “Unpredictability of own future,” and

TABLE 2 Themes of topics derived from free-text responses for academic-based and clinical based researchers

Group	Topic	Theme
Academic-based researchers		
1	1	Stagnation of research support environment
2	2, 4	Unpredictable own future
3	3, 5, 6, 8	Impacts of restrictions
4	7	Barriers to conducting research
Clinical-based researchers		
1	1	Barriers to conducting research
2	2	Increased burden on clinical work
3	3	Barriers to learning/teaching opportunities

“Impacts of restrictions,” whereas the clinical-based researchers produced two additional themes, “Increased burden on clinical work” and “Barriers to teaching/learning opportunities.”

4 | DISCUSSION

This study identified the negative impacts on research activities in the workplace at the outbreak of COVID-19 by investigating free-text responses to a question from a cross-sectional online survey using LDA topic modeling along with PCA. The one theme common in the responses of both academic-based and clinical-based researchers was “Barriers to conducting research.”

Although our results were somewhat predictable from previous literature, this study brought some noteworthy discussions. These were (1) how the generated topics, groups of topics/themes corresponded to the actual free-text responses; (2) how the results were dissimilar to the same question between the pre-defined and free-text responses; and (3) how possibly LDA and PCA could be integrated or incorporated into a process of qualitative data analysis in the future.

Undeniably, the COVID-19 pandemic made it hard for researchers in both groups to conduct their ongoing or new research, especially in the early days. The word cloud of the theme common to both groups visualizes the word “investigation” (調査) as prominently sized, indicating this is the word that most frequently appeared in the relevant free texts (Appendix S1). This result could be explained by the restrictions, during this period, imposed on access to clinical sites by outsiders, including visitors, teaching staff, and students, that were put in place to minimize the spread of the virus. Consequently,

researchers lost opportunities for communication or collaboration for their “investigations”, creating barriers. The original JANS survey report demonstrated that approximately 82% of the members ($n = 899$) felt a negative impact to their research activities (Japan Academy of Nursing Science, 2020). This barrier may also have been faced more by the academic-based researchers than by their clinical counterparts because the theme, “Stagnation of research support environment,” which is close to the common theme in the intertopic distance map, also was observed in responses by academic researchers. They encountered various forms of “stagnation,” including a hiatus in the activity on the part of the Research Ethics Committee or research management department due to the pandemic closure of educational facilities. Measures to mitigate this negative impact on research activities should be focused on providing alternatives means within the research support environment and catering more to the needs of academic-based researchers.

Interestingly, the sets of topic words generated by the LDA for the one common theme were different between the two groups. In particular, the words “month” (月), “child” (子ども), and “time” (時間), which emerged only among the clinical-based researchers, seemed to have different characteristics than other words belonging to the common theme. In the actual text, these words were used to describe the necessity of parenting during school hours due to the nationwide school closures implemented in the early “month[s]” of the pandemic. Accordingly, the clinical-based researchers had to spend “time” on parenting and housework when there was a “child” (or children) at home. This extra time could have impacted their research work. An online survey of 4189 Japanese parents who care for young dependents (aged <12 years) revealed that the mothers more than the fathers significantly increased their time for childcare and housework during the pandemic (Sakuragi et al., 2022). While no significant difference in the demographics of parenting status between the groups was observed in our study, the words noted above could be the pain points for clinical-based researchers who continuously work under challenging circumstances. They perhaps indicate a key point for considerations of countermeasures, such as making the working style and childcare support systems more flexible.

Additionally, our result could be indirectly linked with the results from the aforementioned comparative study by Inoue et al. (2022). They used the same JANS survey as in our study but only examined several pre-defined responses. They reported that several negative factors for research activities were not statistically significant between academic and clinical-based researchers. In other words, these factors negatively affected both groups. Nevertheless, our results which matched those of Inoue et al. (2022) were

found only in one group of either academic- or clinical-based researchers. These were “Unpredictable own future,” “Impacts by restrictions,” and “Barriers to teaching/learning opportunities.” These differences could be based not only on the type of question but also on the respondents’ willingness to comment on the question. Participants tend to avoid answering open-ended questions unless necessary because of inconvenience (Zuell et al., 2015). Yet, open-ended questions often include diverse answers incorporating respondents’ unique perspectives and experience, unlike pre-defined questions (He & Schonlau, 2021; Ozuru et al., 2013). In our study, the theme of “Increased burden on clinical work” was uniquely noted by clinical-based researchers. They have been usually placed in a unique position in various situations such as conducting research, engaging in clinical education, or studying for a higher degree, while caring for patients. This uniqueness may have influenced their research activities. When COVID-19 occurred, clinical-based researchers had to deal with ever-changing measures of infection control, while continuing their primary duties, resulting in increased workload and burden. In contrast, academic-based researchers could have moved to online-based work, except for the restrictions on fieldwork. This change brought additional workload for some people, but it also gave them time for research. While the research environments were fundamentally different between the two groups, our results highlighted the perspectives of clinical-based researchers who are a minority in the JANS and yielded a weighty finding.

The method used in our study may suggest future possibilities for integration or incorporation of topic modeling into qualitative data analysis that requires human coding in nursing research. One of our themes was in close agreement with the findings from a qualitative study using content analysis by Amano et al. (2021). They explored the impediments to nursing research activities during the early stage of the pandemic and reported 12 categories within a framework of five research steps. Although both their study and ours used the same JANS survey data, our study analyzed a single specific open-ended question. Contrary to this, the study by Amano et al. (2021) used all of the open-ended questions in the survey and was written in Japanese. This agreement in our results could be because we followed a process of naming the topics, groups of the topics, and themes to accurately reflect the contexts by reading the actual responses after LDA generated the keywords. The LDA only derived lexical items based on a mathematical model and does not read the nuances in the context or phrases (Guetterman et al., 2018). However, using topic modeling may enhance qualitative data analysis, which is often criticized as being subjective in its results and time-consuming for the processing of data, especially large

amounts of data, due to human coding (Linneberg & Korsgaard, 2019). One study that compared LDA topic modeling with traditional approaches using human coding, including the Heideggerian phenomenological approach, in terms of the results, time, and costs spent on the analysis of the same qualitative textual data, found that topic modeling led to similar results (Abram, 2018; Abram et al., 2020). The same study also showed that the coding process was at least 120 h faster and about \$1500 cheaper with LDA topic modeling compared to the traditional counterpart (Abram et al., 2020). Guetterman et al. (2018) suggested that NLP could be incorporated into qualitative data analysis to validate its findings. These combined methods may increase not only the reproducibility but the reliability of qualitative data analysis by saving time for coding and costs, even with larger sample sizes. Follow-up studies with such combined methods may allow us to explore more deeply the differences in the impacts of COVID-19 between the two groups. However, the method used in this study will become an emerging research area in nursing science.

Several limitations should take into account when drawing conclusions from this study. First, the sample size is relatively smaller, with a low response rate, compared to other studies that use NLP methods. This method usually performs better with larger sample sizes to generate latent topics and themes. Hence the topic and themes we identified may not fully represent the perspectives of the participants. Second, we used LDA, one of the most popular NLP techniques in machine learning. There are many other techniques in this area that may produce different results if the parameters of applied models or the “random seed” values are different. Third, some keywords may have been lost erroneously or assigned nonsensically to a topic because some texts or contents were not clear or had inadequate wording. As a result, some topics or themes often contain various words, which are often tricky for meaningful insights because topic modeling is driven by categorizing words only. Finally, our analysis included the survey participants who only answered the specific open-ended question. Some respondents participated in the survey but were excluded unless they responded to this question in free text. Hence, our results might not be generalizable but still meaningful for contextual understanding.

5 | CONCLUSION

Using LDA topic modeling with PCA, we identified similarities and differences between academic-based and clinical-based researchers in the themes described in free-text responses to a question about the impact of the COVID-19

pandemic on their research activities. Some of the themes had not been identified in an analysis of pre-defined responses to the same survey question. Measures to mitigate the negative impacts of the pandemic on nursing research activities may need to be tailored for academic-based and clinical-based researchers separately. Further, the measures may be better if based on findings gained from analyses of both pre-defined and free-text responses.

AUTHOR CONTRIBUTIONS

Madoka Inoue: Conceptualization, data analysis and interpretation, writing the manuscript. Hiroki Fukahori: Data collection, contribution of scientific knowledge. Manami Matsubara: Conceptualization, contribution of scientific knowledge. Naoki Yoshinaga: Data collection, contribution of scientific knowledge. Hideo Tohira: Conceptualization, data analysis and interpretation, contribution of scientific knowledge. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS


We are extremely thankful to the JANS members who participated in the initial survey during this challenging period. In particular, those who described their opinions, including distress, difficulties, and anxiety, as free-text responses to share with us were much appreciated. Furthermore, yet importantly, our respects are paid to the COVID-19 committee members and the JANS staff for their ongoing support. Open access publishing facilitated by Curtin University, as part of the Wiley - Curtin University agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest.

ORCID

Madoka Inoue  <https://orcid.org/0000-0002-2004-5594>

Naoki Yoshinaga  <https://orcid.org/0000-0002-4438-9746>

REFERENCES

- Abram, M. D. (2018). The role of the registered nurse working in substance use disorder treatment: A hermeneutic study. *Issues in Mental Health Nursing*, 39(6), 490–498. <https://doi.org/10.1080/01612840.2017.1413462>
- Abram, M. D., Mancini, K. T., & Parker, R. D. (2020). Methods to integrate natural language processing into qualitative research. *International Journal of Qualitative Methods*, 19, 160940692098460. [Artn160940692098460810.1177/1609406920984608](https://doi.org/10.1177/1609406920984608).
- Amano, K., Morimoto, H., Watanabe, R., Sato, K., Hiroki, F., Shimpuku, Y., & Yoshinaga, N. (2021). Exploring the obstructive and facilitative factors of nursing research activities during the spread of COVID-19. *Journal of Japan Academy of Nursing Science*, 41, 656–664 (in Japanese).
- Bashri, M., & Kusumaningrum, R. (2017). Sentiment analysis using latent Dirichlet allocation and topic polarity wordcloud visualization. Paper presented at the 2017 5th International Conference on Information and Communication Technology (ICoICT7).
- Beel, J., Gipp, B., Langer, S., & Breitingner, C. (2015). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, 17(4), 305–338. <https://doi.org/10.1007/s00799-015-0156-0>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4–5), 993–1022. <https://doi.org/10.1162/jmlr.2003.3.4-5.993>
- Buenano-Fernandez, D., Gonzalez, M., Gil, D., & Lujan-Mora, S. (2020). Text Mining of Open-Ended Questions in self-assessment of university teachers: An LDA topic modeling approach. *IEEE Access*, 8, 35318–35330. <https://doi.org/10.1109/access.2020.2974983>
- Chung, G., Rodriguez, M., Lanier, P., & Gibbs, D. (2022). Text-mining open-ended survey responses using structural topic modeling: A practical demonstration to understand Parents' coping methods during the COVID-19 pandemic in Singapore. *Journal of Technology in Human Services*, 1–23. <https://doi.org/10.1080/15228835.2022.2036301>
- Friborg, O., & Rosenvinge, J. H. (2011). A comparison of open-ended and closed questions in the prediction of mental health. *Quality & Quantity*, 47(3), 1397–1411. <https://doi.org/10.1007/s11135-011-9597-8>
- Garbhapu, V., & Bodapati, P. (2020). A comparative analysis of latent semantic analysis and latent Dirichlet allocation topic modeling methods using bible data. *Indian Journal of Science and Technology*, 13(44), 4474–4482.
- Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. G. V. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of Medical Internet Research*, 20(6), e231. <https://doi.org/10.2196/jmir.9702>
- He, Z. S. Y., & Schonlau, M. (2021). Coding text answers to open-ended questions: Human coders and statistical learning algorithms make similar mistakes. *Methods Data Analyses*, 15(1), 103–119. <https://doi.org/10.12758/mda.2020.10>
- Inoue, M., Tohira, H., Yoshinaga, N., & Matsubara, M. (2022). Propensity-matched comparisons of factors negatively affecting research activities during the COVID-19 pandemic between nursing researchers working in academic and clinical settings in Japan. *Japan Journal of Nursing Science*, e12491, e12491. <https://doi.org/10.1111/jjns.12491>
- Janome v0.4 documentation (en). (2020). Available from <https://mocabeta.github.io/janome/en/>
- Japan Academy of Nursing Science. (2020). Survey of Members of the Japan Academy of Nursing Science (JANS): Impacts of COVID-19 on Research Activities and Support Expected from JANS. Available from https://www.jans.or.jp/modules/en/index.php?content_id=80#covid19committe
- Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.
- Linneberg, M. S., & Korsgaard, S. (2019). Coding qualitative data: A synthesis guiding the novice. *Qualitative Research Journal*, 19(3), 259–270. <https://doi.org/10.1108/Qrj-12-2018-0012>

- Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). *Optimizing semantic coherence in topic models*. Paper presented at the proceedings of the 2011 conference on empirical methods in natural language processing.
- Ogden, J., & Lo, J. (2012). How meaningful are data from Likert scales? An evaluation of how ratings are made and the role of the response shift in the socially disadvantaged. *Journal of Health Psychology, 17*(3), 350–361. <https://doi.org/10.1177/1359105311417192>
- Onan, A. (2019). Two-stage topic extraction model for bibliometric data analysis based on word Embeddings and clustering. *IEEE Access, 7*, 145614–145633. <https://doi.org/10.1109/ACCESS.2019.2945911>
- Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications, 57*, 232–247.
- Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology, 67*(3), 215–227. <https://doi.org/10.1037/a0032918>
- Pietsch, A.-S., & Lessmann, S. (2019). Topic modeling for analyzing open-ended survey responses. *Journal of Business Analytics, 1*(2), 93–116. <https://doi.org/10.1080/2573234x.2019.1590131>
- Prime Minister of Japan and His Cabinet. (2020). [COVID-19] Press Conference by the Prime Minister (Opening Statement). Available from https://japan.kantei.go.jp/98_abe/statement/202002/_00002.html
- Sakuragi, T., Tanaka, R., Tsuji, M., Tateishi, S., Hino, A., Ogami, A., & CORoNaWorkProject. (2022). Gender differences in housework and childcare among Japanese workers during the COVID-19 pandemic. *Journal of Occupational Health, 64*(1), e12339. <https://doi.org/10.1002/1348-9585.12339>
- Salih Hasan, B. M., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining, 2*(1), 20–30. <https://doi.org/10.30880/jscdm.2021.02.01.003>
- Shaukat, S. S., Rao, T. A., & Khan, M. A. (2016). Impact of sample size on principal component analysis ordination of an environmental data set: Effects on eigenstructure. *Ekológia (Bratislava), 35*(2), 173–190. <https://doi.org/10.1515/eko-2016-0014>
- Verleysen, M., & Francois, D. (2005). The curse of dimensionality in data mining and time series prediction. In J. Cabestany, A. Prieto, & S. Francisco (Eds.), *Computational intelligence and bioinspired systems. 8th international work-conference on artificial neural networks, IWANN 2005 proceedings* (Vol. 3512, pp. 758–770). Springer.
- Vijayan, R. (2021). Teaching and learning during the COVID-19 pandemic: A topic modeling study. *Education Sciences, 11*(7), 347. <https://doi.org/10.3390/educsci11070347>
- Zuell, C., Menold, N., & Korber, S. (2015). The influence of the answer box size on item nonresponse to open-ended questions in a web survey. *Social Science Computer Review, 33*(1), 115–122. <https://doi.org/10.1177/0894439314528091>
- Züll, C. (2016). Open-ended questions (version 2.0). In *GESIS Survey Guidelines*. GESIS – Leibniz Institute for the Social Sciences.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Inoue, M., Fukahori, H., Matsubara, M., Yoshinaga, N., & Tohira, H. (2022). Latent Dirichlet allocation topic modeling of free-text responses exploring the negative impact of the early COVID-19 pandemic on research in nursing. *Japan Journal of Nursing Science*, e12520. <https://doi.org/10.1111/jjns.12520>