



ARTICLE



<https://doi.org/10.1057/s41599-022-01185-6>

OPEN

# Using big data to understand the online ecology of COVID-19 vaccination hesitancy

Shasha Teng <sup>1</sup>✉, Nan Jiang<sup>1</sup> & Kok Wei Khong<sup>1</sup>

With a large population of people vaccinated, it is possible that at-risk people are shielded, and the coronavirus disease is contained. Given the low vaccine uptakes, achieving herd immunity via vaccination campaigns can be challenging. After a literature review, we found a paucity of research studies of vaccine hesitancy from social media settings. This study aims to categorise and create a typology of social media contents and assess the priority of concerns for future public health messaging. With a dataset of 43,203 YouTube comments, we applied text analytics and multiple regression analyses to examine the correlations between vaccine hesitancy factors and vaccination intention. Our major findings are (i) Polarized views on vaccines existed in the social media ecology of public discourse, with a majority of people unwilling to get vaccinated against COVID-19; (ii) Reasons behind vaccine hesitancy included concerns about vaccine safety, potential side-effects, lack of trust in government and pharmaceutical companies; (iii) Political partisan-preferences were exemplified in vaccine decision-making processes; (iv) Anti-vaccine movements with amplified misinformation fuelled vaccine hesitancy and undermined public confidence in COVID-19 vaccines. We suggest public health practitioners engage in social media and craft evidenced-based messages to online communities in a balanced and palatable way.

<sup>1</sup>Faculty of Business and Law, Taylor's University, Subang Jaya, Selangor, Malaysia. ✉email: [shatengsha@gmail.com](mailto:shatengsha@gmail.com)

## Introduction

On 9 November 2020, Pfizer and BioNTech announced an interim vaccine efficacy of >90% for their mRNA vaccine candidate BNT162b2 (Pfizer and BioNTech, 2020). This was followed by Moderna's announcement of their mRNA candidate mRNA-1273 with an efficacy of 94.5% on 16 November 2020 (Moderna, 2020). Interim results from AstraZeneca-Oxford of their viral-vectored ChAdOx1 vaccine were released on 8 December 2020 with an overall vaccine efficacy of 70.4% across two cohorts (Voysey et al., 2021). Pfizer then announced an overall vaccine efficacy of 95% and of 94% in the high-risk group of 65–85 year olds on 10 December 2020 (Polack et al., 2020). These fundamental basic scientific discoveries are expected to pave the way for the successful eradication of the coronavirus disease 2019 (COVID-19). Sputnik V and CanSino vaccines had been used in Russia and China, respectively (Sputnik, 2020; Wee and Qin, 2020). Both the Pfizer-BioNTech and Moderna vaccines were approved for emergency use in the USA. The Pfizer-BioNTech, Moderna, and AstraZeneca-Oxford were authorised for use in the UK. Several other countries approved these vaccines in December 2020 and January 2021 (Carvalho et al., 2021). At the time of writing, 32.5% of the UK population and 38.9% of the US population were fully vaccinated. The data also show that 8.6% of Brazilians and 3.0% of Indians have been vaccinated against COVID-19 (Our World in Data, 2021).

With vaccination programs underway in many countries, it becomes important to examine people's (un)willingness to get vaccinated against COVID-19. Before the COVID-19 vaccine becomes available, a number of survey studies have provided percentages of the populations willing to receive a COVID-19 vaccine, 69% of Americans (Reiter et al., 2020), 69% of British, and 64.9% of Irish (Murphy et al., 2021), 73.9% of Europeans (NeumannBöhme et al., 2020). Research findings also revealed reasons behind vaccine hesitancy and refusal, including concerns about vaccine safety, potential adverse events and side-effects, lack of knowledge of vaccines, lack of trust in pharmaceutical companies (Fisher et al., 2020; Murphy et al., 2021; NeumannBöhme et al., 2020; Reiter et al., 2020). Besides these survey studies, researchers explored social media data (e.g., Twitter tweets, YouTube videos, and Facebook posts) and categorised vaccine sentiments expressed by online users. One study of tweets conveyed that vaccine hesitancy stemmed from the following themes: concerns over safety, suspicion about political or economic forces driving the COVID-19 pandemic, or vaccine development (Griffith et al., 2021). Another study of social media posts categorised 58% of these posts into positive sentiments. Public optimism over vaccine development, effectiveness, and trials were identified (Hussain et al., 2021). As alluded to above, although researchers and regulators stress that the benefits of vaccines outweigh the risks, there is still low vaccine uptake in states and countries. Because of limited socialising during the pandemic, concentrated vaccine information exploded in social media. We thus are motivated to explore vaccine hesitancy factors that are impeding the efforts by public health officials to fight the pandemic. Despite a panoply of previous study findings revealing public perceptions and sentiments toward COVID-19 vaccines from social media data, little attention has been paid to public understandings of vaccine efficacy and their behavioural changes via individual-level social media data over the period of vaccine development and vaccination programs.

Against this background, the objective of this study is to identify and categorise themes of YouTube audience's perceptions of COVID-19 vaccines using big data analytics. We focus on studying real-time data gleaned from YouTube to understand public thoughts, feelings, and behaviours. A thematic analysis of YouTube comments is conducted to identify the YouTube

audience's perceptions. With identified thematic clusters, we then adapt these clusters as constructs to the Health Belief Model (HBM) (Rosenstock et al., 1988). With this predictive model, we undertake a causal modelling approach to explore the correlations and assess the causal relations amongst identified themes. Specifically, the following are investigated: (1) what are the major themes of YouTube audience's discussion of COVID-19 vaccines? (2) how do identified thematic clusters influence the YouTube audience's intention to get vaccinated against COVID-19?

This study provides two key contributions. First, it provides a contribution in terms of research methods that can be implemented by applying text analytics to generate weights from social media data for predictive modelling. We believe it advances the method of big data analytics and generates insights for data analytics researchers. Second, this study adds to previous studies by mapping and prioritising various reasons behind vaccine hesitancy and refusal that can be utilised to devise further vaccine intervention programs. We extended the HBM model by examining existing constructs and newly emerged factors that influence COVID-19 vaccination behaviours. Such insights have practical value for public health professionals and the way they combat the COVID-19 pandemic.

## Literature review

We first cite the definition of vaccine hesitancy and three vaccine hesitancy categories proposed by the World Health Organization (WHO). Specifically, we review the COVID-19 vaccine hesitancy factors identified from several survey studies across the world. We continue to review the literature on the COVID-19 vaccine hesitancy in the social media context. The majority of the reviewed papers focus on sentiment analysis and topics of public discourses. After analysing previous studies, we identify the research gap and propose our research objectives. The comparison of traditional surveys and big data analytics is discussed. We emphasise the relevance and meaningfulness of applying big data analytics to the study of the COVID-19 vaccine hesitancy.

## Vaccine hesitancy in the COVID-19 pandemic

The WHO Strategic Advisory Group of Experts (SAGE) Working Group defines vaccine hesitancy as a delay in acceptance or refusal of vaccination despite the availability of vaccination services. It is complex, context-specific, varying across time, place, and vaccines (MacDonald and the SAGE Working Group on Vaccine Hesitancy, 2015). Dubé and collaborators (2013) have suggested that vaccine hesitancy should be seen on a continuum ranging from active demand for vaccines to complete refusal of all vaccines. As a heterogeneous group, vaccine-hesitant individuals are in the middle of this continuum. These individuals may agree to take some vaccines and refuse others. They may acquiesce in vaccination but be unsure about specific vaccine uptake (Dubé et al., 2013). The Working Group proposes three categories of factors that affect individuals' decision to be vaccinated. These factors include confidence (e.g., trust in vaccine effectiveness and safety, trust in the system and health professionals, and policy-makers motivations), complacency (e.g., perceived risks of vaccine-preventable disease), and convenience (e.g., physical availability, affordability, and willingness to pay, and geographical accessibility). The WHO SAGE Working Group developed the vaccine hesitancy determinants matrix with traditional vaccine hesitancy factors, such as personal experience with vaccination, perceived risks/benefits, geographic barriers, religion, cost, and so on (MacDonald and the SAGE Working Group on Vaccine Hesitancy, 2015). In the context of the ongoing COVID-19 pandemic, vaccine hesitancy factors related to the COVID-19

vaccine may be similar but conceptually distinct from traditional hesitancy factors in the circumstances that vaccine safety is not well-established and not widely available. Callaghan and collaborators (2020) cautioned that reasons for vaccine hesitancy and refusal in previous studies might not be generalised and used for the COVID-19 vaccine studies (Callaghan et al., 2020). It is estimated that the threshold for the population to achieve COVID-19 herd immunity through vaccination or prior infection would be 55% ( $R_0 = 2.2$ ) and 82% ( $R_0 = 5.7$ ) (Sanche et al., 2020). A high vaccine hesitancy rate would significantly affect the attainment of immunity. Consequently, it is crucial to explore and identify the public acceptance of this particular vaccine and their reasons for not pursuing vaccination. Previous studies suggested that COVID-19 vaccination intentions vary substantially across countries (Edwards et al., 2020). Early data from a survey of the United States showed that 31% of participants were unwilling to get a COVID-19 vaccine with a reportedly high level of perceived potential vaccine harm (Reiter et al., 2020). Additional reasons for vaccine hesitancy included vaccine-specific concerns, a need for more information, anti-vaccine attitudes or beliefs, and a lack of trust (Fisher et al., 2020). A nationally representative survey study from the UK reported that 26% of respondents were hesitant about a COVID-19 vaccine (Murphy et al., 2021). Another survey study in seven European countries (i.e., Denmark, France, Germany, Italy, Portugal, the Netherlands, and the UK) found that 18.9% of participants stated they were unsure, and 7.2% said they would not get vaccinated. Reasons for hesitating to get a COVID-19 vaccine included concerns about potential side-effects and vaccine safety (NeumannBöhme et al., 2020). Taken together, vaccine hesitancy and refusal can be major barriers to attaining herd immunity.

### COVID-19 vaccine hesitancy and social media

Social media provides users with non-traditional sources for health information, including blogs and alternative news sites. It offers an outlet to share vaccine opposition beliefs widely in real-time (Deiner et al., 2019). Vast and varied discussions on vaccinations are going on in social media even before the COVID-19 vaccine becomes available. To date, a variety of studies have shown the impact of social media information on public vaccination decision-making. Johnson et al. (2020) provided a system-level analysis of the online ecology of Facebook users' vaccine views. The illustration of the evolution of online ecology showed that anti-vaccination clusters became highly entangled with undecided clusters in the online network. Pro-vaccination clusters were more peripheral in the illustration. The study's theoretical framework predicted that anti-vaccination views would dominate in a decade (Johnson et al., 2020). More recently, Loomba and collaborators (2021) conducted a randomised control trial in the UK and USA to measure the impact of the COVID-19 vaccine misinformation on vaccination intent. The findings indicated that misinformation had induced declines in intent in the UK and USA, with points of 6.2% and 6.4%, respectively. Additionally, scientific-sounding misinformation was associated with reductions in vaccination intents (Loomba et al., 2021). The growing interest in studying the impacts of social media information motivates us to conduct a literature review. Since the review focuses on social media and COVID-19 vaccines, we searched key terms including 'social media,' 'COVID vaccine,' 'Twitter,' 'Facebook,' 'YouTube,' 'Reddit,' 'TikTok,' and 'Instagram,' and a combination of terms in Google Scholar. We limited our search to studies in English. We screened the titles, abstracts, and full texts to evaluate the suitability against inclusion and exclusion criteria. Disagreements were resolved through discussion, and we reached a consensus. Reasons for exclusion of studies were: not addressing

COVID-19 vaccines, contents related to specific vaccines such as influenza, measles, and Human Papillomavirus (HPV), contents related to vaccines in general, not peer-reviewed, and editorials, letters, and commentaries. The resulting articles were coded with reference details, social media platforms, methods, sample size, and main results (see Supplementary Appendix 1). Among fifty-two articles, we identified nine relevant articles related to COVID-19 vaccines in the social media context. Most of the studies applied content analysis of social media data. For instance, Basch and collaborators (2021a) documented the numbers of YouTube videos, numbers of views, and proportion of cumulative views (Basch et al., 2021a). They extended the work to study TikTok videos and explore sentiments of encouraged and discouraged use of COVID-19 vaccines (Basch et al., 2021b). Three studies focused on categorising topics themes, including grouping most reliable, less reliable, and unreliable Twitter information (Jamison et al., 2020), grouping Twitter vaccine hesitancy themes (e.g., safety and political scepticism) (Griffith et al., 2021), and characterising major vaccine hesitancy topics (e.g., conspiracy, developing speed, and safety) (Thelwall et al., 2021). The rest of included articles employed machine learning to investigate sentiments of online data regarding COVID-19 vaccines. One study manually annotated tweets collected after the first vaccine announcement and concluded that most tweets had a neutral stance (Cotfas et al., 2021). Another study analysed the sentiment trends in the UK and US. The geospatial mapping of social media public sentiment helped identify areas with more negative sentiments toward COVID-19 vaccines. The comparative analysis revealed that average Facebook and Twitter comments were mainly positive in both countries (Hussain et al., 2021). Praveen et al. (2021) conducted sentiment analysis and topic modelling to identify Indian citizens' attitudes and concerns about COVID-19 vaccines (Praveen et al., 2021). Despite a number of studies have identified public sentiments and topics of public discourses in major social media platforms, little attention has been paid to explore the impacts of identified topics/factors on social media users' vaccination intention. In this study, we intend to identify public discourse topics themes regarding COVID-19 vaccines. We are motivated to take a further step to explore the relationships between vaccine hesitancy factors and the public's intention to get vaccinated against COVID-19.

### Survey and big data analytics

As we observed the abovementioned acceptance rates of COVID-19 vaccinations across countries, most studies were conducted using surveys. These traditional methods are characterised by slow and costly research, small sample size, and low response rates (On et al., 2019). Traditional survey methods offer useful evidence, both cross-sectional and regional but lack individual-level data. Furthermore, it is difficult to track changes in nuanced beliefs among different populations over time (Dredze et al., 2016). For instance, while using surveys for exploring public concerns about vaccinations, researchers found that the methods were limited in the ability to detect sudden changes in confidence levels (Karafillakis et al., 2017). Yet, it is critical to monitor and analyse changes in the perceptions of vaccinations continuously. In contrast to traditional survey methods, social media analytics allows researchers to observe many individuals' beliefs, especially in the pandemic, without reaching the people directly and physically. Specifically, social media analytics provides unprecedented, real-time access to people's perceptions, attitudes, and behaviours (Dredze et al., 2016). In the case of vaccinations, social media becomes a hotbed for anti-vaccine activities. Applying social media analytics can fill in gaps left by traditional methods on vaccine hesitancy and refusal (Dredze et al., 2016). For

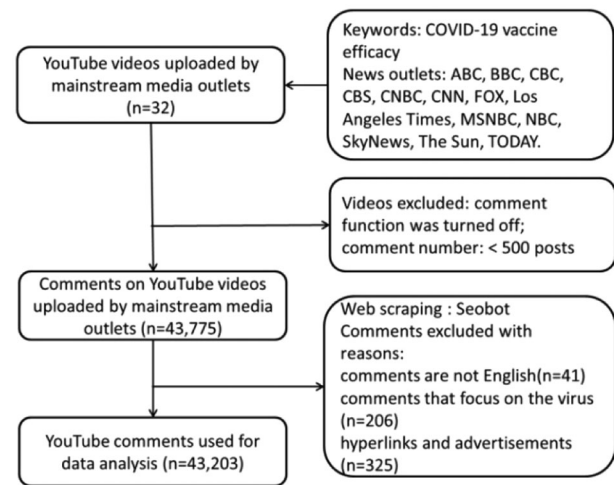
example, Dredze et al. (2016) noted that researchers could better capture the opinions of vaccine refusals and investigate public messages shared by members of a community of interest (Dredze et al., 2016). Moreover, by directly observing online messages, researchers can circumvent bias concerns of participants who respond in ways they are expected during survey data collection. In the case of sensitive and controversial topics, such as vaccine hesitancy and refusal, social media data may provide more accurate information (Dredze et al., 2016). Du et al. (2020) demonstrated the feasibility of deep learning of social media posts in understanding population-level and individual health beliefs and attitudes toward vaccines (Du et al., 2020). Larson (2020) pointed to the need for the use of social media in listening to public concerns in real-time, given the changing nature of vaccine sentiments (Larson, 2020).

In the era of big data, social media data is considered raw and organic materials that need to be purposefully processed, aggregated, and transformed into insights that provide meanings given a specific context (Hallikainen et al., 2020). Insights are attained through analytics, including extracting and interpreting hidden patterns of online public discourses. Previous studies have used social media data for opinion mining, sentiment analysis, and topic modelling to understand health-related issues (Griffith et al., 2021; Hussain et al., 2021). Few studies documented the numbers of YouTube videos, numbers of views, and proportion of cumulative views (Basch et al., 2020). Big data analytics, used in this study, thus refers to acquiring, storing, processing, and analysing a large collection of complex COVID-19 vaccine-related data. Rather than cataloguing YouTube videos, we intend to use the text mining method to dissect YouTube comments, discover information, and advance analyses of the collections of textual data. We use YouTube video viewers' comments to understand their thoughts, feelings, and behaviours. We aim to create meaningful and real-time information to support public health professionals' decision-making for the benefits of increased population vaccination to reduce susceptibility to COVID-19.

## Methods

We conducted mixed-method studies to explore and identify social media discussion patterns and themes and investigate the relationships between identified themes/constructs regarding COVID-19 vaccines. The mix-method approach included a thematic analysis of qualitative data and a causal modelling approach to explore and assess the causal relations using quantitative data derived from the text mining method. A thematic analysis of YouTube video comments with labelled clusters allows us to understand video viewers' prioritised discussion topics in social media conversations. We utilised these clusters to perform multiple regression analyses to examine the relationships between vaccine hesitancy factors and intentions to get vaccinated against COVID-19. This approach also investigates the impacts of misinformation on social media users' hesitancy to take vaccine jabs.

**Data collection.** After reviewing previous studies (Cotfas et al., 2021; Griffith et al., 2021), indicated by Google trends with the highest number of searches, that is, the first announcement of Pfizer-BioNTech vaccine candidate over 90% effective in the first interim efficacy analysis, we focused on the announcements of the COVID-19 vaccine efficacy. We also included the announcements of Moderna and AstraZeneca-Oxford COVID-19 vaccine efficacy in November 2020, which were eligible for our data collection. Thus, the keyword 'COVID-19 vaccine efficacy' was used to search on YouTube. As indicated by previous studies (Basch et al., 2020; Jamison et al., 2020), most YouTube videos and the most reliable topics of the COVID-19 vaccine were



**Fig. 1 The flowchart of data collection process.** It illustrates the design of data collection. Keywords to search relevant videos and video exclusion criteria were established for selecting videos. Once we selected videos, we scraped video comments and excluded specific comments with reasons. A number of 43,203 comments are used for data analysis.

uploaded by news media. We, therefore, limited our selection of video uploaders to accredited mainstream news outlets. Search results were sorted by view count to obtain the most widely viewed YouTube videos in descending order at the time of data collection. Videos that comments function was turned off and comment number was less than 500 posts were excluded. The URLs of each video, video title, and the number of posts were documented in Supplementary Appendix 2. With the list of URLs, we used Botster (ex. Seobot) to scrape YouTube video comments. All collected data were stored in Excel spreadsheets. We removed user nicknames, user URLs, and dates to ensure anonymity and avoid potential bias from data interpretation. Both original comments and replies under original posts were kept for data analysis. We focused on studying YouTube users' comments, and the criteria for the inclusion of user comments are twofold: i. contents are in English, ii. comments that focus on the vaccine instead of the virus itself. We independently screened all comments and removed hyperlinks and advertisements. Figure 1 illustrates the data collection process. A total of 43,203 data entries were returned after we went through the exclusion process. Similar to previous studies' approaches (Basch et al., 2020; Du et al., 2020), this study does not involve human subjects. The data collected are publicly available, and social media user comments are posted on publicly available websites. This study was exempt from the institutional board review.

**SAS text analytics.** We performed text analytics using SAS Text Miner (9.4) with a corpus of large amounts of textual data (documents in SAS Text analytics). The datasets were added as input sources in the abovementioned software. There are four typical processes in text analytics, including input data, text parsing, text filter, and text cluster. The text parsing node in the software was added to extract, clean, and create a dictionary of words from documents (Chakraborty et al., 2013). Each document was divided into tokens (terms). These tokens were listed in the matrix of term-by-frequency, which was used to identify the most frequently occurring words and a number of documents where each word occurred. Meanwhile, the text parsing process detected and determined parts of speech and removed stop words (e.g., and, a, and also) (Chakraborty et al., 2013). Table 1 shows the excerpts of the text parsing results. For instance, some of the

most frequently occurring words are ‘vaccine,’ ‘people,’ ‘take,’ and ‘virus,’ which makes sense as these words are commonly used in discussing vaccine efficacy provided in the scope of this study.

After the text parsing node, the text filter node was added to filter out terms that occurred the least number of times in the collection of documents. Similar terms were formed into groups using synonyms. In other words, grouped terms closely related to each other were characterised by separate clusters of related terms. Thus, terms within one group were similar, and terms between groups were dissimilar (Chakraborty et al., 2013). Term Frequency-Inverse Document Frequency (TF-IDF) algorithm in Supplementary material 1 was used to evaluate how relevant a word is to a document in a collection of documents. Table 2 provides the matrix of documents of terms dropped and kept in the process of text filter.

The text cluster node worked on the basis of identified terms and their frequency of occurrence in the corpus of documents and within each document. At this step, singular value decomposition (SVD) was used to transform the original weighted term-document-matrix into a dense but less dimensional representation (Chakraborty et al., 2013). In other words, the clustering results via SVD on the term frequency document were set to generate enough SVD dimensions (*k*) for further

analysis. The greater the number of dimensions (*k*), the higher the resolution to the term frequency clustered. In this study, *k* was set to 50, which was the default setting in the SAS software. The *k*-dimensional subspace was then generated from the cluster analysis via SVD. The cluster frequency root mean square standard deviation shown in Table 3 indicates that the derived clusters via SVD were well manifested and optimal as the values are close to 0. Regarding the number of clusters, we made some trial-and-error in the properties setting of text cluster in an attempt to generate well-separated clusters. A 40-cluster solution using a hierarchical clustering algorithm was set in the text cluster node. The results showed a set of descriptive terms of each cluster in Table 3, which revealed to a certain extent of the theme of each cluster.

Figure 2 presents the cluster hierarchy graph. The distance between the clusters depicts the relationships between these clusters. The further the clusters are from each other; the less likely the key themes between the clusters are associated. In short, the distance between clusters manifests the association amongst the key themes. The endpoints of hierarchical clusters are sets of subclusters (e.g., 19 Pfizer and 21 Trump in Fig. 2), where each subcluster is distinct from other subclusters. Regarding clusters, descriptive terms in each cluster (e.g., Cluster 20 in Table 3) are broadly similar to each other, evidenced by shared terms (Chakraborty et al., 2013). In this study, for instance, *democrats won* and *right* related to the term *political ideologies*. The cluster *perceived trust in pharma* was mainly associated with *companies* and *billion*.

**Results**

The text analytics we performed in the methods section produced eleven different clusters. These clusters are reflected in the most frequent key terms used by the YouTube video commenters. We used an inductive approach to interpret each cluster’s theme. This is accomplished by individually reviewing sentences extracted from YouTube video comments. As suggested by Glowacki et al. (2017), we compared our assessments, and disagreements were resolved by group discussions to arrive at this final description. We then explained each cluster with excerpts of YouTube comments. We interpreted data clusters with a narrative approach to

**Table 1 Excerpts of the text parsing results.**

TERM	rolestring	attrstring	Freq	numdocs	keep
+ be	Verb	Alpha	30837	17212	Drop
not	Adv	Alpha	19451	13085	Drop
+ vaccine	Noun	Alpha	12409	9386	Keep
+ do	Verb	Alpha	12531	9081	Drop
+ have	Verb	Alpha	10670	7631	Drop
+ get	Verb	Alpha	5807	4667	Drop
+ people	Noun	Alpha	5838	4529	Keep
+ take	Verb	Alpha	5095	4338	Drop
no	Adv	Alpha	4305	3672	Drop
+ say	Verb	Alpha	4085	3372	Drop

Note: + depicts the parent term. Alpha: if characters are all letters. Keep: Keep if the term is used in the text mining analysis.

**Table 2 Excerpts of the text filter results.**

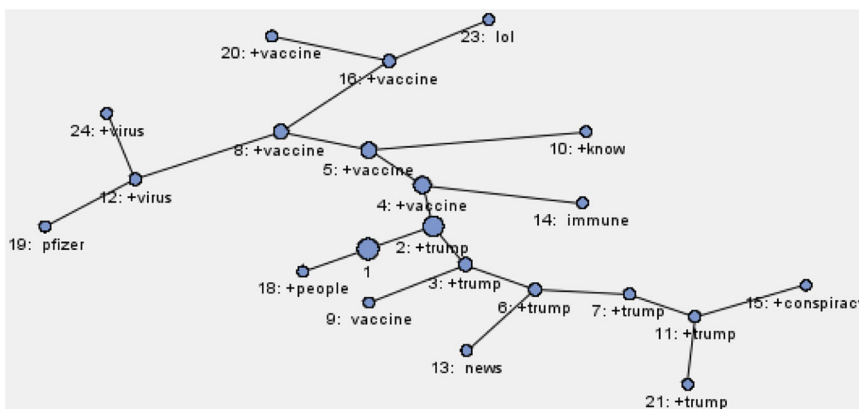
TERM	rolestring	attrstring	Status	Weight	Freq	numdocs	Rank
+ be	Verb	Alpha	Drop	0	30837	17212	1
not	Adv	Alpha	Drop	0	19451	13085	2
+ vaccine	Noun	Alpha	Keep	0.154	12409	9386	3
+ do	Verb	Alpha	Drop	0	12531	9081	4
+ have	Verb	Alpha	Drop	0	10670	7631	5
s	Noun	Alpha	Drop	0	8190	6218	6
+ get	Verb	Alpha	Drop	0	5807	4667	7
+ people	Noun	Alpha	Keep	0.222	5838	4529	8
+ take	Verb	Alpha	Drop	0	5095	4338	9
no	Adv	Alpha	Drop	0	4305	3672	10
+ say	Verb	Alpha	Drop	0	4085	3372	11
+ go	Verb	Alpha	Drop	0	3539	3059	12
+ make	Verb	Alpha	Drop	0	3583	3052	13
just	Adv	Alpha	Drop	0	3052	2774	14
+ know	Verb	Alpha	Keep	0.263	3209	2766	15
+ virus	Noun	Alpha	Keep	0.271	3537	2692	16
what	Adv	Alpha	Drop	0	2682	2490	17
+ trump	Noun	Alpha	Keep	0.273	2857	2487	18
+ year	Noun	Alpha	Keep	0.285	2684	2230	19
+ think	Verb	Alpha	Drop	0	2356	2137	20

Note: + depicts the parent term. Alpha: if characters are all letters. Keep: Keep if the term is used in the text analytics. Weight: value assigned by SAS server.

**Table 3 Text cluster results.**

Cluster ID	Descriptive terms	Freq.	Percent	RMSSTD
20	+vaccine vaccines effective years +flu +year covid pfizer months first +trust developed ready next +end	5750	13%	0.122045
21	+trump biden +president trump +credit +election joe +big +media democrats won +right america china donald	4636	11%	0.128844
19	pfizer +government +money +company +want german biontech + 'german company' companies germany gates bill warp +billion speed	4296	10%	0.119179
10	+know +election +right +day days +mask knew knows masks gonna dr +wear wearing convenient	4081	9%	0.130781
14	immune +system +long + 'immune system' +time +media +term +body effects +keep +truth bbc +side + 'long term' +watch	3909	9%	0.136214
18	+people effects gates 'side effects' +god side bill +die +side died +wake +care +million +want vaccine	3981	9%	0.122634
23	lol +want +good +comment +thing +science things +time yes comments yeah well +read dont +hope	3916	9%	0.121946
24	+virus +world +people +mark viruses +beast +cure +rate +corona +body +cold coronavirus +population common +pandemic	3838	9%	0.129316
9	vaccine first effective covid-19 +test cases trials testing data +million +population tested numbers deaths +number	3697	8%	0.134275
13	news +good trump +rate survival +'survival rate' fake president better +death 'good news' 'fake news' fox +great luck	3444	8%	0.128687
15	+conspiracy theories +evidence +theory facts 'conspiracy theories' theorists +comment true +fact 'conspiracy theorists' +'conspiracy theory' +idiot +theorist +moron	2067	5%	0.120634

Note: + depicts parent term. Cluster ID: Hierarchical clustering IDs do not range from 1 to 10. Freq: number of times that the term appears in the document collection. Percent: percentage of documents in a cluster. 'immune system' and 'side effects' are noun phrases that grouped together based on linguistic relationships in the unigram text parsing process. RMSSTD root mean-square standard deviation.



**Fig. 2 Cluster hierarchy graph.** It presents cluster hierarchy in the study. Clusters are represented as nodes. The distance of two clusters is modelled by a certain weight on the edge linking the nodes.

understand meaningful human experiences (see Supplementary material 2). We labelled each cluster based on the most frequently used terms and their weights calculated by SAS Text Miner. Table 4 shows the summary of cluster interpretation in descending order of cluster frequency indicated in the data analysis.

Traditionally, identifying and grouping themes requires several rounds of coding, adding, and consolidating codes in order to reach a point of adequate saturation. This approach is based on code appearances across the corpora. Consistent with calculating term frequency but distinctively different from the traditional coding method, this study adopted text clusters derived from a predetermined algorithm embedded into SAS Text Miner, which was deemed accurate and reduced bias compared to human coding. We then conducted supervised learning with these identified clusters by independently reading YouTube video comments and labelling these clusters. In line with the traditional survey item designs, we labelled clusters based on measurable items in statistical analysis, such as vaccination intention and perceived trust. We first read descriptive terms of each cluster and carried out an initial reading of about 20% of comments ( $n = 8000$ ) independently among researchers. This step allows us to understand a narrative trope of our data regarding the

COVID-19 vaccines. At a later stage, we carried out reading of the remaining 80% of comments ( $n = 35,203$ ) independently. The only manually coding process in this study was labelling identified clusters. Based on the literature review and comments reading in the earlier stages, we managed to label each cluster by continuously corroborating the insights from the dataset. We also constantly triangulated the data with previous studies and the results of this study through virtual discussions due to the lockdown. One of the researchers conducted a quality check to ensure the assumptions and preconceptions of other researchers had not biased the labelling.

The criteria of labelling clusters are using the most frequently occurring descriptive terms produced from the text cluster algorithm. Regarding Cluster 20, the majority of YouTube video audience were not willing to get vaccinated. The likelihood of taking recommended preventive health action is an essential means of measuring health-related behaviours in the Health Belief Model (HBM) (Janz and Becker, 1984; Rosenstock et al., 1988). In this study, we thus labelled Cluster 20 as intention to get COVID-19 vaccines. Cluster 21 was identified as political ideologies. Previous study findings revealed that Republican (Red) states had significantly lower unadjusted HPV vaccine series

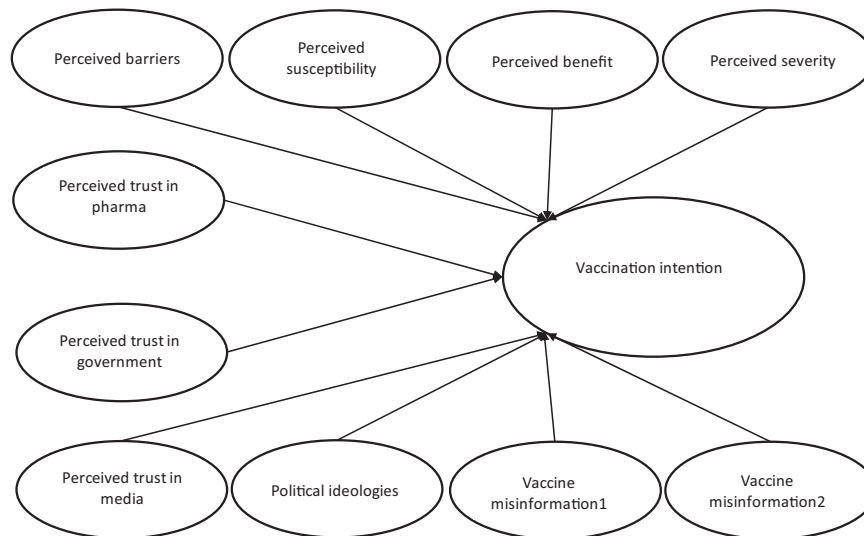
**Table 4 Summary of text cluster interpretation.**

Cluster label/percent	Descriptive terms	Cluster summary	Cluster quotes
Cluster ID 20 Vaccination intention/ 13%	+vaccine vaccines effective years +flu +year covid pfizer months first +trust developed ready next +end	The dominant coronavirus vaccine narratives are the unwillingness to take a vaccine, discussions of reasons for refusing to take a jab, questions on the vaccine efficacy, and concerns about vaccine side-effects.	<i>I definitely will NEVER take an mRNA vaccine. "Up to 90%" only means that it is less than 90%, I suppose it kills the other 10% +. I'll give it a miss thanks all the same, the virus only has a 2% fatality rate Democrats you promised to not take the Trump vaccine. Can we take you for your word on that.</i>
Cluster ID 21 Political ideologies/11%	+trump biden +president trump +credit +election joe +big +media democrats won +right america china donald	Most of the online conversations about the US election revolved around Trump and Biden. The majority of YouTube viewers want the politicians to take the vaccine first so that they can wait and see the vaccine effects.	<i>It's a very good business to sell stuff to governments that they then force on the people. That's why big pharma companies are striving to be first. And, of course, "booster shots" will be needed annually. NO WAY I'M TAKING A VACCINE!! AFTER tRUMP I DON'T TRUST GOVERNMENT</i>
Cluster ID 19 Perceived trust in pharma/10%	pfizer +government +money +company +want german biotech + 'german company' companies germany gates bill warp +billion speed	Discussions of Pfizer-BioNTech vaccine centred on the Operation Warp Speed. Criticism of big pharma included it making profits for getting everyone vaccinated when vaccine mandates advanced.	<i>The MSM will hide any side effects, unless Trump prevails, then they will make up horror stories of death, mutilation etc.</i>
Cluster ID 10 Perceived trust in government/9%	+know +election covid +right +day days +mask knew knows masks gonna dr +wear wearing convenient	People questioned the political motives of the timing of broadcasting the COVID-19 vaccine efficacy. People distrust the government and tend to disregard government-mandated disease-control measures.	<i>I won't be taking any vaccine provided. I'm expecting side effects.</i>
Cluster ID 14 Perceived trust in media/9%	immune +system +long + 'immune system' +time +media +term +body effects +keep +truth bbc +side + 'long term' +watch	There were negative narratives and conversations surrounding the mainstream media, including biased, parroting, and fear-mongering reports. YouTube users stated that media is the enemy of people.	<i>Free 5 G Microchips. Contact Billy Boy - If the tests don't infect you, then the Vaccine will make you a Transhuman Slave.</i>
Cluster ID 18 Perceived barriers/9%	+people effects gates 'side effects' +god side bill +die +side died +wake +care +million +want vaccine	A great number of YouTube viewers warned the video audience not to take the COVID-19 vaccine. The majority of the conversation contents are rushed vaccine, unknown adverse effects, untested and unlicensed vaccine.	<i>Without care, 1 in 5 people who catch the virus die. That's over a 20% mortality rate. the 99.97% is how much of society has both been unexposed and survived exposure, so it's very misleading. 90% effective vs the covid survival rate of 99% hm soooo 10% chance the vaccine won't work, and a +-0.1% of dying from Covid? Sounds like a lot of people will get sick from this vaccine, or any for that matter. (sm)</i>
Cluster ID 23 Vaccine misinformation/9%	lol +want +good +comment +thing +science things +time yes comments yeah well +read dont +hope	There was a spike in conspiracy theory contents related to the coronavirus vaccine. Many questioned their source credibility.	<i>No need to rush the vaccine—2/3 of the country won't take it anyway!!!</i>
Cluster ID 24 Perceived severity/9%	+virus +world +people +mark viruses +beast +cure +rate +corona +body +cold coronavirus +population common +pandemic	People believe the mortality rate of the COVID-19 pandemic is 2%, which is very low compared to SARS and MERS.	<i>I'm not getting the vaccine #pandemic</i>
Cluster ID 9 Perceived benefit/8%	vaccine first effective covid-19 +test cases trials testing data +million +population tested numbers deaths +number	The online conversations centred on the sample size (43,000 subjects), ethnic diversity, and placebo test subjects. Some believe the vaccine phase 3 trial data is preliminary and not peer-reviewed, indicating that people have doubts over the vaccine's effectiveness.	
Cluster ID 13 Perceived susceptibility/8%	news +good trump +rate survival + 'survival rate' fake president better +death 'good news' 'fake news' fox +great luck	YouTube users believe that they are at a low risk of getting infected. Conversations around the COVID-19 vaccine refusal were also connected to the survival rate of infected people (i.e., 99.8%). People also think the survival rate is higher than the COVID-19 vaccine efficacy. They do not think they need to take the vaccine.	
Cluster ID 15 Vaccine misinformation/5%	+conspiracy theories +evidence +theory facts 'conspiracy theories' theorists +comment true +fact 'conspiracy theorists' +'conspiracy theory' +idiot +theorist +moran	The majority of the content about conspiracy theories is related to the coronavirus vaccine. Key terms included '5 G,' 'microchip,' 'agenda 21,' 'New World Order,' and 'China released the virus.'	

The italic texts are quotes (posts) from YouTube videos comment section. Regardless of grammatical errors, we intend to keep their original posts for the purpose of illustrating examples of related clusters. Although we included a series of anecdotes illustrating some rumours, we actually intend to invite the public to debunk the myths. We critically examined our role of supporting vaccination and the researcher reflexivity guides us through this research study.

initiation and completion rates (Suryadevara et al., 2019). A recent survey study has found that participants were willing to get vaccinated if they were moderate or liberal in their political leaning (Reiter et al., 2020). We included political ideologies as one important factor in influencing preventive health behaviours. A content analysis of anti-vaccine websites has shown a great deal of mistrust in multiple entities (e.g., government, pharmaceutical companies, and healthcare providers) expressed by online users.

The values and beliefs of lying and irresponsible entities contributed to the distrust of vaccine advocate organisations (Moran et al., 2016). One recent survey study examined the impact of trust in pharmaceutical companies on COVID-19 vaccination acceptance (Wong et al., 2021). We thus included perceived trust in pharma as one crucial contextual factor of health-related behaviours. Cluster 10 was labelled as perceived trust in government. We included this cluster as another contextual factor of



**Fig. 3 Research model.** The proposed research model with several independent variables and one dependent variable.

health-related behaviours. Some YouTube video viewers believe that the Mainstream Media (MSM) is lying and fear-mongering. As indicated by Moran et al.'s (2016) study findings, mistrust in mass media is one important contextual factor in predicting vaccination against vaccine-preventable diseases. We thus labelled Cluster 14 as perceived trust in media.

According to Rosenstock et al. (1998), perceived barriers refer to negative components of an anticipated behaviour, including financial costs, physical barriers, side-effects, and accessibility factors (Rosenstock et al., 1988). Perceived barriers, such as vaccine safety concerns and costs, were significant predictors of vaccine uptake (Dubé et al., 2013; Gerend and Shepherd, 2012). Cluster 18, labelled as perceived barriers, was used to measure the impact of vaccine safety concerns on vaccine hesitancy in this study. We identified Cluster 23 as vaccine misinformation. We observed that misinformation regarding COVID-19 vaccines was closely related to anti-vaccine conspiracy theories. In contrast to anti-vaccine groups, some YouTube users urged others to stop spreading misinformation and get vaccinated to protect people and save lives. Perceived severity refers to a concern about how threatening the condition is to the person. In the case of vaccines, previous studies have proved that perceived severity was strongly associated with HPV vaccine uptake (Gerend and Shepherd, 2012), and willingness to get vaccinated against COVID-19 (Reiter et al., 2020). We thus hypothesised that the perceived severity of COVID-19 infection would be an important factor in influencing hesitancy. Given the descriptive terms generated from unsupervised machine learning, we confirmed that the keyword effectiveness reflected people's perceived benefit of the vaccine during the supervised learning. Thus, we labelled cluster 9 perceived benefit. Perceived benefit focuses on the belief regarding the effectiveness of a specific behaviour or alternative behaviour in preventing disease, maintaining health, and lessening undesirable consequences of the disease (Rosenstock et al., 1988). Perceived benefit, for instance, safety and effectiveness of a vaccine, accounted for the major variations in seeking vaccines. Gerend and Shepherd (2012) examined the relationships between HBM constructs (e.g., perceived benefit) and HPV vaccine acceptability (Gerend and Shepherd, 2012). Positive correlations of perceived benefit and COVID-19 vaccine acceptance were also found in a survey study (Wong et al., 2021). We adapted the perceived benefit to this study, intending to evaluate its impact on

vaccination intention. We identified and labelled Cluster 13 as perceived susceptibility. It refers to a person's view of the likelihood of experiencing a potentially harmful condition. Reiter et al. (2020) found that participants were likely to be willing to get vaccinated if they reported a higher level of perceived likelihood of getting a COVID-19 infection in the future. Perceived susceptibility was also measured in Sherman et al.'s (2020) study of public attitudes and beliefs in COVID-19 vaccination acceptability (Sherman et al., 2020). We adapted perceived susceptibility to examine its impact on COVID-19 vaccination intention.

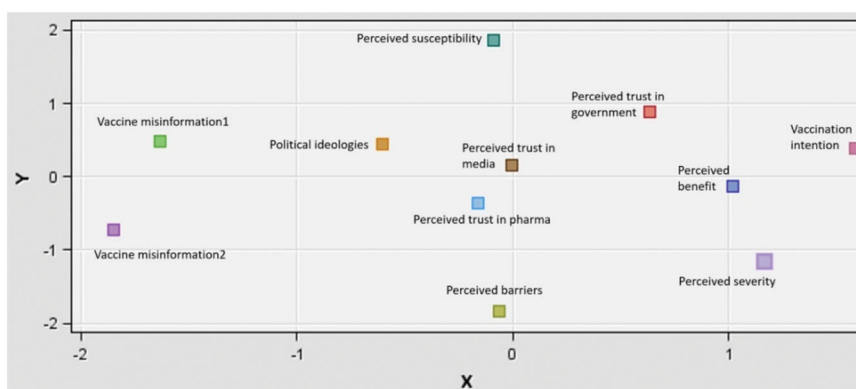
Cluster 15 was labelled as vaccine misinformation. Based on our review of the COVID-19 vaccine hesitancy and social media, previous studies have found negative associations between misinformation and vaccine uptake intention. For instance, misinformation induced the decline in vaccine uptake (Loomba et al., 2021), inaccurate and negative social media data potentially influenced population-wide vaccine uptake (Basch et al., 2021a), belief in COVID-related conspiracy theories predicted resistance to preventive health behaviour and vaccination against COVID-19 (Romer and Jamieson, 2020). We intend to adapt vaccine misinformation as a contextual factor in our research framework. Moreover, we explore and measure HBM constructs and contextual factors (e.g., perceived trust in pharmaceutical companies, vaccine misinformation). We expect to observe vaccine hesitancy-induced behavioural change in the COVID-19 pandemic. Figure 3 is the proposed research model that depicts the relationships of constructs in the context of the COVID-19 vaccines. Table 5 shows the constructs of the research model.

**Reliability and validity of unsupervised and supervised learning.** With the help of unsupervised machine learning, we conducted data collection and analysis, which was thoroughly and scientifically analysed. The process of understanding data refers to the initial exploration and evaluation of the collected data. Given the specific phenomenon (i.e., the COVID-19 vaccine efficacy) and study objectives, we decided to use YouTube comments generated from mainstream media outlet videos. As noted in the methods section, the cleaning data process needed to be done manually. We verified that the data collected was accurate, confirming that comments were related to the COVID-19 vaccine. This process of preparing data was thus found to be highly reliable. At the data analysis stage, we emphasised the choice and



**Table 5 Constructs of the study model.**

Cluster ID	Cluster construct	Cluster description
20	Vaccination intention	Audience are unwilling to get vaccinated against COVID-19.
21	Political ideologies	Conservatives or Republicans are more likely to reject a COVID-19 vaccine compared to Liberals or Democrats.
19	Perceived trust in pharma	The public distrusts pharmaceutical companies who make profits from vaccines, and they are exempted from legal liabilities.
10	Perceived trust in government	The public distrusts the government and politicians. People questioned the political motives of withholding the announcement of Pfizer vaccine efficacy.
14	Perceived trust in media	YouTube users stated that mainstream media pushed lies and propaganda to the public.
18	Perceived barriers	The public raised concerns surrounding potential harms regarding rushed COVID-19 vaccines.
24	Perceived severity	People believe the mortality rate of the COVID-19 pandemic is relatively low, and the condition is not threatening to them.
9	Perceived benefit	YouTube users believe COVID-19 vaccines would not prevent future problems for them.
13	Perceived susceptibility	People believe that it is less likely they contract COVID-19 disease. They also think the survival rate is higher than the COVID-19 vaccine efficacy.
23&15	Vaccine misinformation	A great number of anti-vaccine individuals and groups shared conspiracy theories. Concerns were raised about the spread of misinformation.



**Fig. 4 Cartesian plot of cluster distances.** The identified clusters were plotted according to the distance generated from a cluster algorithm.

calibration of the methods to analyse data. We intended to validate the results compared to the objectives of the study. A hierarchical process was preferred in the text cluster algorithm (Fig. 2). We chose the combination of SVD and optimal *k*-means as it was judged to be a good way of balancing unstructured data at hand with the purpose of the study (Carnerud, 2017). The root mean square standard deviation (Table 3) showed the average distance between observations in clusters. It indicated the measure of goodness of fit of text cluster. At the stage of cluster interpretation, we consistently labelled each cluster on the basis of the frequency of occurrence of texts as labelling is an approved, preferred, and effective process of operating qualitative data (Carnerud, 2017). The supervised learning of cluster interpretation consisted of detailed notes and codes, accurate interpretation of comments, data transparency and authenticity, critical appraisal of cluster labelling. We tested validity through the convergence of information from various sources (e.g., news, previous studies). We are able to develop a comprehensive understanding of the social media ecology of the COVID-19 vaccine through the use of triangulation of qualitative research. This human judgement allowed us to validate, evaluate, and interpret generated clusters consistently and efficiently, leading to an achieved reliability and validity of this study.

**Multiple regression analyses.** We observe that several clusters are related to each other, making it interesting to explore them using an appropriate method. With derived clusters, we use a Cartesian coordinate system by displaying them in a scatterplot (see Fig. 4).

The Cartesian coordinate measures the distance between clusters with X and Y coordinates representing space generated by the Singular Value Decomposition (SVD) algorithm. Clusters aligned together in the Cartesian system represent their similarities in semantic concepts. The values derived from text cluster (see Supplementary material 3) are the mean values. These values measure the distance of a data point to the centroid in each cluster. Given our study’s text cluster parameter setting, the iterative algorithm produced 46 ‘samples’. We believe that any changes of the distances of data points to one centroid in one cluster in the iterations would impact the distances to another centroid in another cluster. Using these values, we developed a predictive model where the independent variables such as perceived susceptibility, perceived barriers, perceived benefit, perceived trust in pharma, political ideologies, and vaccine misinformation were associated with vaccination intention. Unlike traditional survey items, there was single item/variable corresponding to each ‘sample’ for each cluster in this study (*n* = 46). Given the context of values generated from text clusters, we adopted observed variables representing each construct (Table 5). There are one dependent variable (i.e., vaccination intention) and ten independent variables in the proposed research model. We used multiple regression analysis to examine the impacts of the abovementioned variables on vaccination intention in this study. Multiple regression analysis is one of the most used statistical techniques to analyse the relationship between a single dependent variable and several independent variables. This regression analysis aims to evaluate the proposed research model by determining the influence of the independent variables on the single dependent variable. Given the fact that

**Table 6 Multiple regression analyses for the COVID-19 vaccination intention predictors.**

Model summary <sup>b</sup>						
Model	R	R square	Adjusted R square	F	Std. error of the estimate	
1	0.864 <sup>a</sup>	0.747	0.692	13.628	0.03340223	
ANOVA <sup>b</sup>						
Model	Sum of squares	df	Mean square	F	Sig.	
1	Regression Residual Total	8 37 45	0.122 0.041 0.163	0.015 0.001	13.628	
Coefficients <sup>b</sup>						
Model	Unstandardised coefficients		Standardised coefficients		Sig.	Collinearity statistics
	B	Std. error	Beta	t		
1	Tolerance					
	(Constant)	0.000	0.005	0.013	0.990	
	Ideologies	0.401	0.150	2.663	0.011	0.353
	Pharma	0.131	0.107	1.222	0.230	0.546
	Media	0.139	0.116	1.200	0.238	0.324
	Barriers	0.159	0.098	1.617	0.114	0.541
	Susceptibility	0.350	0.117	2.981	0.005	0.424
	Benefit	-0.141	0.181	-0.777	0.442	0.249
	Severity	-0.094	0.130	-0.723	0.474	0.254
	Misinformation2	0.058	0.119	0.484	0.631	0.450

a. Predictors: (constant), misinformation2, pharma, media, barriers, susceptibility, ideologies, severity, benefit.

b. Dependent variable: intention.

generated values were obtained from text clusters and proposed research model was hypothesised from the HBM literature and cluster labelling, multiple regression analysis is a clear viable solution to develop predictive modelling in relation to vaccination intention. The analyses were carried out using SPSS (22.0).

First, we ran the SPSS analysis to identify outliers. We followed the rule of thumb to take a boxplot approach. Supplementary material 4 is the boxplot of all clusters. It is obvious that case 1 is the outlier. After dropping the outlier, the R square value fell over 40%. The coefficients indicated that dropping this outlier did not significantly change the results or affect assumptions. Thus, we kept this outlier for further analysis.

Second, we performed linear regression analysis with all identified variables. We observed the variance inflation factor (VIF) of collinearity statistics. In a regular multiple regression, the path coefficients may be biased if the estimation involves significant levels of collinearity among the predictor constructs (Hair et al., 2010; Hair et al., 2014). Hair et al. (2014) considered VIF above 5.00 as indicative of collinearity that is too high. Hence, we removed one independent variable at a time based on the VIF values in descending order. For instance, vaccine misinformation1 was dropped, followed by perceived trust in government.

Third, after checking for collinearity, we continued to assess the nonlinearity and homogeneity of variance. The bivariate plot of the predicted value against residuals can infer whether the predictors' relationships to the outcome are linear. The scatterplot (supplementary material 5) presents the standardised predicted value of independent variables on the dependent variable against the standardised residuals. From the Loess curve, it appears that the relationship of standardised predicted value to residuals is roughly linear around zero. It does not suggest a departure from linearity, and we would expect the Loess line to come close to the regression line. Moreover, the residuals seem to be randomly scattered around zero, we thus conclude that the linearity assumption is satisfied, and the heteroskedasticity assumption is satisfied if we run the fully specified predictive model (Cohen et al., 2003; Keith, 2019).

Fourth, we tested the normality of residuals using a normal probability plot (more specifically, a P-P plot). The P-P plot (supplementary material 6) compares the observed cumulative distribution function (CDF) of the standardised residual to the expected CDF of the normal distribution. In addition, the quantile-quantile (q-q) plot (supplementary material 7) compares the observed quantile with the theoretical quantile of a normal distribution. As can be seen from the Q-Q plot graph, the residuals conform fairly well to the diagonal straight line. We conclude that the residuals are normally distributed (Cohen et al., 2003).

Fifth, we performed multiple linear regression analyses to elucidate whether identified clusters/variables (e.g., political ideologies, perceived susceptibility, or perceived trust in pharma) or an interaction of these variables were predictors of vaccination intention among social media users. We focused on the total effects of variables (in particular, the prediction purpose) on an outcome (Keith, 2019). The results of multiple regression analyses (see Table 6) suggested two predictor variables were significantly related to vaccination intention. The  $R^2$  of 0.747 indicates that 74.7% of the variance can be accounted for by its linear relationship with the predictor variables. An  $R^2$  of 0.700 is described as close to substantial (Hair et al., 2010). It indicates that the regression model fits the observed data well. Adjusted  $R^2 = 0.692$ ,  $F(8, 37) = 13.628$ ,  $p < 0.001$ . According to the standardised weight  $\beta$  in the following table, the regression equation is as follows:

$$\text{Vaccination intention} = 0.379_{\text{susceptibility}} + 0.371_{\text{ideologies}} + 0.182_{\text{barriers}} + 0.174_{\text{media}} + 0.137_{\text{pharma}} + 0.060_{\text{misinformation2}} - 0.119_{\text{severity}} - 0.129_{\text{benefit}}$$

The coefficients matrix provides information about the relative influence of independent variables. It showed that perceived susceptibility statistically influenced vaccination intention ( $\beta = 0.379$ ,  $t = 2.981$ ,  $p = 0.005$ ). In other words, the perceived likelihood of getting infected with COVID-19 could influence their intention to get vaccinated. In addition, political ideologies is related to vaccination intention ( $\beta = 0.371$ ,  $t = 2.663$ ,  $p = 0.011$ ). It indicates that social media users' political ideologies could influence their intentions to get vaccinated against COVID-19. The results also showed that perceived trust in pharma, perceived trust in media, perceived barriers, perceived benefit, perceived severity, and vaccine misinformation2 were not significantly associated with vaccination intention. To this end, our cluster interpretation and predictive model generate meaningful results that provide us with better understandings of vaccine hesitancy factors and HBM constructs in influencing COVID-19 vaccination intention.

## Discussion

In previous sections, we interpreted and labelled eleven clusters generated from SAS Text analytics. Based on the thematic analysis, we adapted vaccine hesitancy factors and other contextual factors into the HBM constructs in an attempt to develop a research model regarding intention to get COVID-19 vaccines. We investigated a variety of casual relationships among the constructs by employing the multiple regression analysis method. Now we will discuss our major findings of COVID-19 vaccine hesitancy factors and public perceptions and behaviour changes. We will include theoretical and practical implications in this section as well.

**Major findings.** Using individual-level social media data of 43,203 YouTube comments, we reveal a number of important findings to policy-makers and stakeholders engaged in the public health sector. First, we find that the social media ecology of vaccine hesitancy is evolving. Polarized opinions and contrasting insights characterise the public discourses in this ecology. The profusion of true and false information provides us with a granular understanding of vaccine hesitancy in the social media context, especially amid the COVID-19 pandemic. Apart from vaccine opponents and proponents, a sizable proportion of YouTube users hesitated to get vaccinated against COVID-19. Consistent with previous studies (Basch et al., 2021a; Wong et al., 2021), their reasons included concerns about side-effects, effectiveness, and lack of trust in corporations and government. In addition, this study's thematic analysis findings conveyed that public's perceived trust in media influenced public's vaccination intention. We thus argue that big data analytics of social media posts is a useful tool to aptly generate extensive research results for healthcare professionals to monitor and advance their strategies in improving COVID-19 vaccine uptake.

Second, this study identified reasons behind vaccine hesitancy toward COVID-19 vaccines among YouTube audience, including concerns about vaccine safety, effectiveness, lack of knowledge of COVID-19, and distrust of companies. We categorised labelled clusters into constructs: perceived susceptibility, perceived barriers, perceived benefit, perceived severity, perceived trust in pharmaceutical companies, perceived trust in government, perceived trust in media, political ideologies, vaccine misinformation, and vaccination intention. Perceived trust in companies, government, media, and political ideologies are newly emerged themes compared to previously reviewed articles (Basch et al., 2021a; Hussain et al., 2021). We also observed and identified several health belief model constructs. Similar to previous survey

findings (Gerend and Shepherd, 2012; Wong et al., 2021), perceived susceptibility, perceived barriers, and perceived severity played important roles in the public's decision-making process. Prior studies have measured the impact of vaccine misinformation on the population's vaccine uptake (Loomba et al., 2021; Romer and Jamieson, 2020). Similar to these findings, an additional theme, vaccine misinformation, was also presented in this text analytics study.

Third, based on the multiple regression analyses results, we found the relative importance of health belief model constructs and vaccine hesitancy contextual factors. The top significant factors influencing vaccination intention included perceived susceptibility and political ideologies. This ranking of factors allows healthcare professionals to facilitate target vaccination campaigns. The results showed that perceived susceptibility and political ideologies were strongly associated with COVID-19 vaccination intention. In other words, we can use HBM constructs to explain and estimate COVID-19 vaccine uptake behaviours. In particular, perceived susceptibility could influence social media users' willingness to get vaccinated. The result of perceived susceptibility influence is aligned with previous studies (Gerend and Shepherd, 2012; Romer and Jamieson, 2020) but contrasts with Wong et al.'s (2021) study. The explanation stated that the COVID-19 was considered a mild disease (Wong et al., 2021). Furthermore, our current analysis suggested that social media users have a partisan preference with regard to vaccine information promulgated on YouTube despite the quality of sources. These users tend to choose sources that reflect a similar ideology to their personal beliefs (Suryadevara et al., 2019). One survey study found no evidence that political ideologies were directly related to vaccination intention (Romer and Jamieson, 2020). However, as evidenced by the regression results, YouTube audience whose political leaning toward conservatives are less likely to get vaccinated. Moreover, the right-wing has more tendencies to reduce physical distancing (Gollwitzer et al., 2020), promote vaccine misinformation, and these echo chambers amplify anti-vaccine beliefs in social media platforms (Larson, 2020; Thelwall et al., 2021). Additionally, social media users identified as political conservatives expressed less degree of trust in the government and experts. The vaccine issue is deeply ingrained in politics. Verger and Dubé (2020) noted that vaccine distrust was enmeshed in social and political protest and criticism of vaccines became the hobby horse of opposition parties (Verger and Dubé, 2020). This is particularly troubling because the government and its agencies are the key disseminators of vaccine information. Hence, maintaining and restoring public trust is a necessary component of successful health communication strategies (Moran et al., 2016). Similar to previous studies' findings (Loomba et al., 2021; Romer and Jamieson, 2020), our results cannot draw causal links between perceived trust in media and vaccination intention. However, the data interpretation indicated that mainstream media YouTube channels may be affecting the public responses to vaccine uptake.

**Theoretical and practical implications.** This study makes important contributions to the research methods and theory. Our method of applying text analytics via frequencies and weights of words using social media data is an important contribution. With user-generated contents (texts) transformed into a set of numbers, we are able to use data mining algorithms to generate insights for predictive modelling (Chakraborty et al., 2013). We believe this is a unique combination of big data text analytics and predictive modelling in the research of vaccine hesitancy. Another key contribution of this study is to empirically demonstrate the relevance of HBM constructs and vaccine hesitancy factors in the context of COVID-19 vaccines. Chiming with theoretical constructs, perceived susceptibility, perceived barriers, perceived severity, and perceived

benefit were tested in previous survey studies but have not been tested on a large scale using real-time data. The evidence of perceived susceptibility and political ideologies having effects on vaccination intention validated these factors in the pandemic. The examination of vaccine misinformation as part of vaccine hesitancy factors is an extension of the HBM constructs. Furthermore, this study provides an extensive view of the social media ecology of COVID-19 vaccines. With individual-level vaccine perceptions illuminating pathways, we believe our research contributes to our understandings of existing beliefs that influence how people receive information, not only political-related information but also healthcare-related information.

This study has implications for practice. We provide necessary knowledge about vaccine hesitancy factors that played roles in the public's decision-making of COVID-19 vaccine uptake, including political ideologies. As evidenced by previous studies in France (The COCONEL Group, 2020) and Europe (Kennedy, 2019), positive connections were found between political beliefs and attitudes toward vaccines. In the case of COVID-19 vaccines, political parties and politicians were found to suppress science, denigrate health experts, and peddle disinformation (Abbasi, 2020; Fidler, 2020). For instance, former US president Trump manipulated the Food and Drug Administration (FDA) to approve drugs hastily (Abbasi, 2020). Some expressed concerns that political appointees may insist on the Emergency Use Authorisation vaccine over the recommendation of FDA scientists (Bauchner et al., 2020). Moreover, the UK's pandemic response relies heavily on scientists and other government appointees with worrying competing interests (Abbasi, 2020). Such interferences from governments would present a risk to the public and cause further erosion of public trust in the government's ability to make critical scientific decisions (Bauchner et al., 2020). It leads to dwindling vaccine uptakes in the long run. The best approach for government's pandemic responses in exceptional times is the independence of science agencies and politicians kept from the pronouncements of vaccine-related issues, which can create the appearance of interferences (Lurie et al., 2020). Abbasi (2020) has proposed two steps to safeguard science from politicisation: full disclosure of competing interests from government, politicians, scientific advisors, and appointees; and full transparency of decision-making systems and processes. Politicians have a higher responsibility to the public (Abbasi, 2020). Hence, a root-and-branch reconstruction of political interests on infectious diseases is required in developing and rebuilding health policies sufficiently (Fidler, 2020).

Given a strengthening anti-vaccine movement and continued politicisation of vaccines, public health organisations can utilise this study's methods and findings of vaccine hesitancy factors to facilitate target intervention strategies of combating vaccine hesitancy and vaccine misinformation. First, practitioners are encouraged to take a creative approach to communicate with the public on social media platforms instead of traditional framed messages, authoritative and fact-filled, in promoting vaccines. For instance, online health communication should provide harm and benefit information. Personal narratives such as storytelling style are found to harbour public interests and reinforce public trust from experts or relatable peers (Thomas and Pollard, 2020). A key action for public health professionals is to provide statistical strength of evidence in an anecdotal and palatable way in social media accounts. It is critical to construct a trustworthy information environment that engaging in evidence-based health communication strategies (Thomas and Pollard, 2020). Moreover, practitioners are advised to take vaccination promotions out of medical settings and outreach to social media platforms where the public obtain and share health-related information. They should be cognizant of the negative effects of anti-vaccine on undecided population segments in vaccine uptake.

Additional efforts are needed to keep track of vaccine sentiment and take pre-emptive measures to mitigate any drawbacks. Second, rather than rebutting conspiracy theories, Ahmed (2021) believed practitioners should inoculate against this vaccine misinformation and join local communities in the Facebook groups to offer answers to community members about COVID-19 vaccines (Ahmed, 2021). Meanwhile, online users should resist falling into the trap of engaging in vaccine misinformation and learn to share good information from trusted sources. To effectively address misinformation and public concerns, Thomas and Pollard (2020) recommend public health organisations and practitioners adapt to this digital era and utilise social media platforms to point out logical flaws or malicious intent of anti-vaccine arguments. This online presence of immunisation dialogue is useful in dispelling misconceptions, addressing concerns, and improving vaccination coverage. Lastly, social media companies in 2019 have pledged to act against the anti-vaccine movement, such as Facebook announced that it would not recommend misinformation content, and YouTube removed advertisements from anti-vaccine videos. Ahmed (2021) offers a simple and sterner solution to combat anti-vaccine misinformation, to remove anti-vaccine misinformation super spreaders. Another study recommends social media firms provide visible rewards for users who generate reliable materials, such as adding a 'trust button' and displaying the number of 'trust clicks' as posts received (Sharot, 2021). To this end, we believe that strategic intervention initiatives by public health organisations and social media companies can counteract the spread of COVID-19 vaccine misinformation.

**Limitations and future studies.** This study is delimited in scope in the following ways. First, the data collected was limited to the period of announcements of COVID-19 candidate vaccine efficacy in November 2020. This period coincided with the 2020 US election. We thus expected that a kerfuffle over political views and controversial political figures. In addition, future studies can be conducted to observe longitudinal behavioural changes with different pivotal points in the vaccine development process and vaccination programs. Second, the sampling bias may exist due to the selected social media data. Future studies are encouraged to collect a greater amount of social media data from multiple platforms in an attempt to reduce method bias. Meanwhile, we believe understanding current social media users' attitudes would play a critical role in predicting future generations' vaccination intention. Third, we did not have the data on YouTube users' demographic information while preserving users' anonymity during the data collection. We thus cannot analyse the correlates of demographic characteristics (e.g., age, gender, and ethnicity) to their vaccine-hesitant behaviours. Fourth, as our approach of using text analytics in predictive modelling in vaccination intentions is relatively novel, future studies can be conducted to develop survey scales for vaccine hesitancy factors in an attempt to assess emerging contextual factors and validate the proposed predictive modelling. With these survey scales, we encourage researchers to collect a larger survey sample size to reduce factor dimensions, investigate effect size ( $f^2$ ), and explore reasons for unsupported statistical relationships in the current study. We applied a parsing and stemming process using the tokenization based on the SAS dictionary-based stemmer. To reduce probabilities of cluster overlapping, it is advisable for researchers to train and analyse textual data using an n-gram algorithm, where the computation of linguistics involving two or more terms is possible. In particular, this approach will require a more complex algorithm and a large amount of computational power and resources. Lastly, cues to action construct was not included in identified themes because our independent review of all data found no evidence of this often neglected HBM construct. This factor refers to the signals and reminders from doctors and nurses

to individuals to engage in health-related behaviour. We suggest obtaining relevant data in future studies to examine the effects of cues to action on influencing public vaccine hesitancy and vaccine uptake via personal health beliefs.

## Conclusions

In summary, this study identified the YouTube audience's perceptions of COVID-19 vaccines. Our data interpretation provides major reasons why YouTube audience may feel hesitant or refuse to receive a COVID-19 vaccine. These reasons fell under the following themes: perceived susceptibility, perceived severity, perceived benefit, perceived barriers, perceived trust in pharmaceutical companies, government, media, political ideologies, and vaccine misinformation. We offer these findings hoping that public health practitioners can implement more effective public health communication strategies and tailored interventions in the social media platforms. A better understanding of various social, political, and psychological factors discussed in the study can be used to maximise the positive effects of public health communications. The engagement and participation of key stakeholders are required to convince a sufficient proportion of individuals to get vaccinated against COVID-19, thereby achieving the threshold necessary for herd immunity.

## Data availability

The data that support the findings of this study are available at <https://doi.org/10.7910/DVN/CKBBB5>.

Received: 3 June 2021; Accepted: 26 April 2022;

Published online: 06 May 2022

## References

- Abbasi K (2020) Covid-19: politicisation, "corruption," and suppression of science. *BMJ* 37(1):m4425
- Ahmed I (2021) Dismantling the anti-vaxx industry. *Nat Med* 27(1):363–366
- Basch CE et al. (2021a) YouTube videos and informed decision-making about COVID-19 vaccination: successive sampling study. *JMIR Public Health Surveill* 7(5):e28352
- Basch CH, Hillyer GC, Zagnit EA, Basch CE (2020) YouTube coverage of COVID-19 vaccine development: implications for awareness and uptake. *Hum Vaccine Immunother* 16(11):2582–2585
- Basch CH et al. (2021b) A global pandemic in the time of viral memes: COVID-19 vaccine misinformation and disinformation on TikTok. *Hum Vaccine Immunother*. 17(8):2373–2377
- Bauchner H, Malani PN, Sharfstein J (2020) Reassuring the public and clinical community about the scientific review and approval of a COVID-19 vaccine. *JAMA* 324(13):1296–1297
- Callaghan T et al. (2020) Correlates and disparities of intention to vaccinate against COVID-19. *Soc Sci Med* 272(1):113638
- Carnerud D (2017) Exploring research on quality and reliability management through text mining methodology. *Int J Qual Reliab Manag* 34(7):975–1014
- Carvalho T, Krammer F, Iwasaki A (2021) The first 12 months of COVID-19: a timeline of immunological insights. *Nat Rev Immunol* 21(4):245–256
- Chakraborty G, Pagolu M, Garla S (2013) Text mining and analysis: practical methods, examples, and case studies using SAS. SAS Institute, Cary
- Cohen J, Cohen P, West SG, Aiken LS (2003) Applied multiple regression/correlation analysis for the behavioral sciences, 3rd ed. Routledge, New York, NY
- Cotfas L-A et al. (2021) The longest month: analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *IEEE Access* 9(1):33203–33223
- Deiner MS et al. (2019) Facebook and twitter vaccine sentiment in response to measles outbreaks. *Health Inform J* 25(3):1116–1132
- Dredze M, Broniatowski DA, Smith M, Hiyard KM (2016) Understanding vaccine refusal: why we need social media now. *Am J Prevent Med* 50(4):550–552
- Dubé E et al. (2013) Vaccine hesitancy: an overview. *Hum Vaccine Immunother* 9(8):763–1773
- Du J et al. (2020) Use of deep learning to analyze social media discussions about the human papillomavirus vaccine. *JAMA Network Open* 3(11):e2022025

- Edwards B, Biddle N, Gray M, Sallis K (2020) COVID-19 vaccine hesitancy and resistance: Correlates in a nationally representative longitudinal survey of the Australian population. *PLoS ONE* 16(3):e0248892
- Fidler DP (2020) Vaccine nationalism's politics. *Science* 369(6505):749
- Fisher KA et al. (2020) Attitudes toward a potential SARS-CoV-2 vaccine: a survey of U.S. adults. *Ann Intern Med* 173(12):964–973
- Gerend MA, Shepherd JE (2012) Predicting human papillomavirus vaccine uptake in young adult women: comparing the health belief model and theory of planned behavior. *Ann Behav Med* 44(2):171–180
- Glowacki EM, Glowacki JB, Wilcox GB (2017) A text-mining analysis of the public's reactions to the opioid crisis. *Subst Abuse* 39(2):129–133
- Gollwitzer A et al. (2020) Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nat Hum Behav* 4:1186–1197
- Griffith J, Marani H, Monkman H (2021) COVID-19 vaccine hesitancy in canada: content analysis of tweets using the theoretical domains framework. *J Med Internet Res* 23(4):e26874
- Hair JF, Black WC, Babin BJ, Anderson RE (2010) *Multivariate data analysis*, 7th edn. Prentice-Hall, Upper Saddle River, New Jersey
- Hair JF, Hult GTM, Ringle C, Sarstedt M (2014) *A primer on partial least squares structural equation modeling (PLS-SEM)*. SAGE, Thousand Oaks
- Hallikainen H, Savimäki E, Laukkanen T (2020) Fostering B2B sales with customer big data analytics. *Ind Mark Manag* 86(1):90–98
- Hussain A et al. (2021) Artificial intelligence-enabled analysis of public attitudes on facebook and twitter toward COVID-19 vaccines in the United Kingdom and the United States: observational study. *J Med Internet Res* 23(4):e26627
- Jamison AM et al. (2020) Not just conspiracy theories: Vaccine opponents and proponents add to the COVID-19 'infodemic' on Twitter. *The Harvard Kennedy School Misinformation Review* 1(3):1–22.
- Janz NK, Becker MH (1984) The health belief model: a decade later. *Health Educ Quart* 11(1):1–47
- Johnson NF et al. (2020) The online competition between pro- and anti-vaccination views. *Nature* 582(1):230–234
- Karafilakis E, Larson HJ, On behalf of the ADVANCE consortium (2017) The benefit of the doubt or doubts over benefits? A systematic literature review of perceived risks of vaccines in European populations. *Vaccine* 35(1):4840–4850
- Keith TZ (2019) *Multiple regression and beyond: an introduction to multiple regression and structural equation modeling*, 3rd edn. Routledge, New York, NY
- Kennedy J (2019) Populist politics and vaccine hesitancy in Western Europe: an analysis of national-level data. *Eur J Public Health* 29(3):512–516
- Larson HJ (2020) *Stuck: how vaccine rumors start and why they don't go away*, 1st edn. Oxford University Press, New York, NY
- Loomba S et al. (2021) Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nat Hum Behav* 5(1):337–348
- Lurie N, Sharfstein JM, Goodman JL (2020) The development of COVID-19 vaccines: safeguards needed. *JAMA* 324(5):439–440
- MacDonald NE, the SAGE Working Group on Vaccine Hesitancy (2015) Vaccine hesitancy: definition, scope and determinants. *Vaccine* 33(1):4161–4164
- Moderna (2020) Moderna's COVID-19 vaccine candidate meets its primary efficacy endpoint in the first interim analysis of the phase 3 COVE Study. [Online] Available at: <https://investors.modernatx.com/news-releases/news-release-details/modernas-covid-19-vaccine-candidate-meets-its-primary-efficacy>
- Moran MB et al. (2016) What makes anti-vaccine websites persuasive? A content analysis of techniques used by anti-vaccine websites to engender anti-vaccine sentiment. *J Commun Healthcare* 9(3):151–163
- Murphy J et al. (2021) Psychological characteristics associated with COVID-19 vaccine hesitancy and resistance in Ireland and the United Kingdom. *Nat Commun* 12(29):1–15
- Neumann-Böhme S et al. (2020) Once we have it, will we use it? A European survey on willingness to be vaccinated against COVID-19. *Eur J Health Econ* 21(1):977–982
- On J, Park H-A, Song T-M (2019) Sentiment analysis of social media on childhood vaccination: development of an ontology. *J Med Internet Res* 21(6):e13456
- Our World in Data (2021) Coronavirus Pandemic (COVID-19). [Online] Available at: <https://ourworldindata.org/coronavirus>
- Pfizer & BioNTech (2020) Pfizer and BioNTech announce vaccine candidate against COVID-19 achieved success in first interim analysis from phase 3 study. [Online] Available at: <https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-announce-vaccine-candidate-against>
- Polack FP et al. (2020) Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med* 383(27):2603–2615
- Praveen S, Ittamalla R, Deepak G (2021) Analyzing the attitude of Indian citizens towards COVID-19 vaccine-A text analytics study. *Diab Metab Syndrome: Clin Res Rev* 15(1):595–599
- Reiter PL, Pennell ML, Katz ML (2020) Acceptability of a COVID-19 vaccine among adults in the United States: how many people would get vaccinated? *Vaccine* 38(1):6500–6507
- Romer D, Jamieson KH (2020) Conspiracy theories as barriers to controlling the spread of COVID-19 in the U.S. *Soc Sci Med* 263(1):113356
- Rosenstock IM, Strecher VJ, Becker MH (1988) Social learning theory and the health belief model. *Health Educ Behav* 15(2):175–183
- Sanche S et al. (2020) High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerg Infect Dis* 26(7):1470–1477
- Sharot T (2021) To quell misinformation, use carrots— not just sticks. *Nature* 591(7850):347
- Sherman SM et al. (2020) COVID-19 vaccination intention in the UK: results from the COVID-19 vaccination acceptability study (CoVAccS), a nationally representative cross-sectional survey. *Hum Vaccine Immunother* 17(6):1612–1621
- Sputnik (2020) Second interim analysis of clinical trial data showed a 91.4% efficacy for the Sputnik V vaccine on day 28 after the first dose; vaccine efficacy is over 95% 42 days after the first dose. [Online] Available at: <https://sputnikvaccine.com/newsroom/pressreleases/second-interim-analysis-of-clinical-trial-data-showed-a-91-4-efficacy-for-the-sputnik-v-vaccine-on-d/>
- Suryadevara M et al. (2019) Associations between population based voting trends during the 2016 US presidential election and adolescent vaccination rates. *Vaccine* 37(1):1160–1167
- The COCONEL Group (2020) A future vaccination campaign against COVID-19 at risk of vaccine hesitancy and politicisation. *Lancet* 20(7):769–770
- Thelwall M, Kousha K, Thelwall S (2021) Covid-19 vaccine hesitancy on English language Twitter. *Profesional de la información* 30(2):e300212
- Thomas TM, Pollard AJ (2020) Vaccine communication in a digital society. *Nat Mater* 19(1):476
- Verger P, Dubé E (2020) Restoring confidence in vaccines in the COVID-19 era. *Exp Rev Vaccine* 19(11):991–993
- Voysey M et al. (2021) Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet* 397(10269):99–111
- Wee S-L, Qin A (2020) China approves Covid-19 vaccine as it moves to inoculate millions. [Online] Available at: <https://www.nytimes.com/2020/12/30/business/china-vaccine.html>
- Wong MC et al. (2021) Acceptance of the COVID-19 vaccine based on the health belief model: a population-based survey in Hong Kong. *Vaccine* 39(1):1148–1156

## Acknowledgements

This work is funded by the Ministry of Education Malaysia and Taylor's University (FRGS/1/2019/SS09/TAYLOR/01/1).

## Author contributions

KK and ST were involved in formulating research questions and designing the study. ST and KK performed data management and data analyses. All authors were involved in the interpretation of the data analysis. ST drafted the manuscript. All authors provided input on revised manuscript and approved the final manuscript. All authors agreed to be accounted for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

## Competing interests

The authors declare no competing interests.

## Ethical approval

This article does not contain any studies with human participants performed by any of the authors. The study was deemed exempt from Human Ethics Committee approval by the author's university as it uses publicly available documents or information. This is in compliance with the University's policy on Human Ethics.

## Informed consent

This article does not contain any studies with human participants performed by any of the authors.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1057/s41599-022-01185-6>.

**Correspondence** and requests for materials should be addressed to Shasha Teng.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022