PLOS ONE

RESEARCH ARTICLE

Understanding the determinants of vaccine hesitancy in the United States: A comparison of social surveys and social media

Kuleen Sasse[1], Ron Mahabir[2], Olga Gkountouna[2]*, Andrew Crooks[3], Arie Croitoru[4]

1 Department of Computer Science, The Johns Hopkins University, Baltimore, Maryland, United States of America, **2** Geographic Data Science Lab, Department of Geography and Planning, University of Liverpool, Liverpool, United Kingdom, **3** Department of Geography, University at Buffalo, Buffalo, New York, United States of America, **4** Department of Computational and Data Sciences, George Mason University, Fairfax, Virginia, United States of America

* Olga.Gkountouna@liverpool.ac.uk

Citation: Sasse K, Mahabir R, Gkountouna O, Crooks A, Croitoru A (2024) Understanding the determinants of vaccine hesitancy in the United States: A comparison of social surveys and social media. PLoS ONE 19(6): e0301488. https://doi.org/10.1371/journal.pone.0301488

Editor: Keumseok Peter Koh, University of Hong Kong, HONG KONG

Received: October 15, 2023

Accepted: March 12, 2024

Published: June 6, 2024

Data Availability Statement: All relevant data are within the manuscript.

Funding: The author(s) received no specific funding for this work;.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The COVID-19 pandemic prompted governments worldwide to implement a range of containment measures, including mass gathering restrictions, social distancing, and school closures. Despite these efforts, vaccines continue to be the safest and most effective means of combating such viruses. Yet, vaccine hesitancy persists, posing a significant public health concern, particularly with the emergence of new COVID-19 variants. To effectively address this issue, timely data is crucial for understanding the various factors contributing to vaccine hesitancy. While previous research has largely relied on traditional surveys for this information, recent sources of data, such as social media, have gained attention. However, the potential of social media data as a reliable proxy for information on population hesitancy, especially when compared with survey data, remains underexplored. This paper aims to bridge this gap. Our approach uses social, demographic, and economic data to predict vaccine hesitancy levels in the ten most populous US metropolitan areas. We employ machine learning algorithms to compare a set of baseline models that contain only these variables with models that incorporate survey data and social media data separately. Our results show that XGBoost algorithm consistently outperforms Random Forest and Linear Regression, with marginal differences between Random Forest and XGBoost. This was especially the case with models that incorporate survey or social media data, thus highlighting the promise of the latter data as a complementary information source. Results also reveal variations in influential variables across the five hesitancy classes, such as age, ethnicity, occupation, and political inclination. Further, the application of models to different MSAs yields mixed results, emphasizing the uniqueness of communities and the need for complementary data approaches. In summary, this study underscores social media data's potential for understanding vaccine hesitancy, emphasizes the importance of tailoring interventions to specific communities, and suggests the value of combining different data sources.

Introduction

The impact of the COVID-19 pandemic has resulted in governments worldwide having to implement a slew of different measures to contain the spread of the disease. Such measures, for example, included canceling mass gathering activities, mandating social distancing, school closures, and travel restrictions [1]. However, while these efforts have had some efficacy in slowing disease spread, vaccinations still remain the safest, most effective, and viable approach [2, 3]. As of January 2024, more than 13.5 billion doses of a COVID-19 vaccine had been administered globally. This has resulted in about 71% of the world population having received at least one dose of a vaccine, and 65% being fully vaccinated [4]. Previously, adults 80 years or older were found to be more predisposed to COVID-19; following which, there was increased susceptibility in young adults (aged 18–24 years), and among children and adolescents (aged 0–17 years) [5]. Now, several variants later, with new variants on the rise [6] and with recent COVID-19 outbreaks reported [7], it's critical more than ever to vaccinate in order to continue slowing the disease's transmission. This is vital so that those within the population who cannot be vaccinated, including the very young and immunocompromised, are still protected [8]. Further, while achieving herd immunity may not be feasible due to the evolving nature of the virus [9, 10], ongoing vaccination efforts remain essential to mitigating its impact and safeguarding public health.

Globally, while the total number of people that have received a COVID-19 vaccine has improved over time, these numbers vary by country. In the US, for example, only about 67% of the US population have been fully vaccinated [11]. However, at the state level, these numbers vary, with the lowest and highest vaccination rates being 52.8% (i.e., Wyoming) and 92.2% (i.e., District of Columbia) respectfully [12], and with even further discrepancy at more disaggregated spatial levels. Such variations in vaccination rates within and amongst countries have prompted investigations of the underlying factors that lead people to delay acceptance, or refuse vaccines despite their availability, a phenomenon referred to as vaccine hesitancy [13]. Factors that have been associated with vaccine hesitancy include ethnicity, working status, religious beliefs, political views, gender, age, education, income [14], online misinformation [15], and specific moral values [16]. Yet the majority of such work has depended on conventional surveys to collect data from individuals or groups using questionnaires or interviews that can be in-person (e.g., [17]), online (e.g., [18]), or over the phone (e.g., [19]). While these data sources have contributed to a large and diverse knowledge base on vaccine hesitancy, it's important to consider their limitations (e.g., various biases or sample size) which may limit the applicability and effectiveness of the results (discussed further in the next section).

More recently, newer sources of data, in particular, social media, has emerged as a promising source of information on vaccine hesitancy. Currently, about 60% of the global population (i.e., 4.9 billion people) use social media services, such as X (formerly known as Twitter and henceforth used interchangeably) and Facebook, with this number expected to increase to 5.9 billion people by 2027 [20]. During the COVID pandemic, many people turned to social media as a way of keeping connected and informed about the pandemic [21]. Twitter, in particular, saw a 10.3% increase in users from 2019 to 2020 (the peak of pandemic) [22]. Other social media platforms, such as Facebook, also saw a notable increase of 8.7% [23]. The large number of users on social media makes it a rich source of information on peoples' opinion and sentiments towards various topics, including the pandemic, providing valuable insights into public attitudes and trends [24, 25].

Although there there exists a substantial body of research that have utilized social media data to examine various aspects of vaccines, much of the research have focused only on a few themes. These include analyzing misinformation campaigns and particular communities, such

as the anti-vaccination movement (e.g., [26, 27]), exploring the network interactions among hesitant community members (e.g., [28, 29]), understanding sentiments towards vaccines (e.g., [30, 31]) and the role of social media in influencing public attitude towards them (e.g., [32–34]), and analyzing topics of discourse surrounding vaccines (e.g., [35, 36]). However, the potential of social media data as a viable proxy for understanding vaccine hesitancy and its underlying determinants, particularly when juxtaposed against traditional survey data, remains an area that has garnered relatively little attention. Such an investigation is of significant importance given the profound and ongoing impacts of COVID-19 on our society, and the increasing influence of social media in our interconnected digital age. Further, given that pandemics are expected to continue to occur [37], it is important to continue to identify opportunities for collecting reliable information at scale, reasonable cost, and in a timely manner to help inform public health strategies and interventions. To address this gap, the primary research objective of this study is to evaluate the utility of social media data as a proxy measure for understanding vaccine hesitancy. Additionally, we aim to explore the determinants of vaccine hesitancy using modern machine learning approaches across a broader geographical scope compared to previous work.

Related work

Although research on vaccine hesitancy is not new [38, 39], the advent of the recent COVID-19 pandemic has sparked a massive resurgence of interest surrounding this topic. In particular there has been a growing interest in understanding the specific reasons behind individuals' reluctance to accept the COVID-19 vaccine. Work by [14], for example, delved into the pertinent literature and identified several key factors contributing to vaccine hesitancy. These include apprehension arising from the expedited development of vaccines, the perception of minimal risk regarding the disease due to prior immunization, skepticism surrounding the origin and efficacy of existing vaccines, and a pervasive lack of confidence in the institutions responsible for their production and distribution. Similar findings have been reported by [18, 40–42] with respect to individual determinants of vaccine hesitancy. Other research have found that a lack of time to get a vaccine [43], distrust in the political entities advocating for vaccinations [44], infringements on individual autonomy regarding vaccine accessibility at some locations, conspiracy theories [45], and commercial profiteering [46], further result in increase hesitancy rates. Studies have additionally reported hesitant populations within more specialized groups such as medical professionals (e.g., doctors and nurses) [47], and parents that have not cared for positive COVID-19 cases [48]. Moreover, work by [49, 50] found that even when provided with scientific information to support the efficacy and safety of vaccines, some parents still opt to not vaccinate their children. This suggests that much broader social and cognitive processes may be at play when it comes to making a decision on whether or not to vaccinate [51].

Most of the aforementioned studies have relied on the use of traditional survey instruments, through in–person meetings or online surveys, to elicit insights into vaccine hesitancy. However, these approaches come with inherent limitations that pose challenges in effectively gathering data about vaccines and understanding individuals' reservations towards them. As it relates to in–person surveys; these are very labor intensive, time consuming, and require substantial financial resources [52]. Such costs can be as much as $40 per-person for in-person surveys, or $22 per person for more cost effective options, such as the use of mobile phones using interactive voice response [53], and with data collection costs increasing [54]. This makes it difficult, or at least very costly, to up-scale such work to large geographical scales.

Another concern that has been recognized is that of biases. In the acquiescence bias, survey respondents exhibit a tendency to favor positive response options or express a positive sentiment in a disproportionately frequent manner [55]. Previous work by [56] using China and the US as case studies, for example, showed that acquiescence bias can inflate estimated incidence of conspiratorial beliefs and political perceptions by as much as 50%. Related to this is the dissent bias where people tend to express a negative agreement in a more frequent manner [57]. Survey results can also be affected by social desirability bias wherein respondents choose responses that they believe will make them be viewed favorably by others [57].

In addition to these concerns, traditional surveys face a range of challenges related to participation and timeliness. Declining response rates over time [58], can lead to issues with unit non-response, where participants do not respond to all questions or do not provide enough information for the response to be deemed usable. This is further compounded by the issue of item non-response, in which participants respond to questions but do not provide a usable response to a particular item or items [59]. Beyond participation, surveys can suffer from problems of temporal relevancy, which represents the need to have data collected as close as possible in time to the event of interest [60, 61]. As such, they must be planned ahead of time, which would typically mean having knowledge about an event that is either yet to happen, or the ability to collect participants' information on short notice following an event. For mass emergency and rapidly unfolding events, such as natural disasters and epidemics, advanced prior knowledge may be limited, and participants may otherwise be pre-occupied during these times to take part in surveys. Further, because surveys typically represent the current view of participants as of the date of the survey, they are unable to adequately address issues from bias stemming from experiences that may have occurred prior to the administered survey date [60]. For example, data collected from people after they have had severe side effects from a vaccine may lead to a negative view towards vaccines, and could influence other members within their social circle to not vaccinate. Collectively, these biases can convolute public opinion, potentially leading to distorted interpretations of the true beliefs and perceptions landscape.

In light of the shortcomings of traditional surveys, other sources of data have been explored. One such source is social media data collected from online platforms such as Twitter and Facebook. Compared to survey data, social media presents several advantages, including the ability to scale data collection efforts at reasonable cost, permit archival searches to capture more temporal relevant data surrounding events of interest, as well as other relevant information that were not included in the original survey instrument [60]. With respect to the latter, whereas surveys tend to be restricted to what is required, thus leading to a higher possibility of issues with unit and item non-response, the content on social media remains largely unbounded. This, in turn, can increase the potential for capturing additional useful information about the particular event or phenomenon that can be of value. These benefits are also expected to extend to instances of acquiescence and dissent bias as well, with social media allowing for more freedom of expression [62], compared to the use of poorly constructed survey instruments that typically lead to such issues [63].

Prior research on surveys has demonstrated that respondents experience lower social anxiety and social desirability when participating in online surveys as opposed to face–to–face interactions, which can be attributed to the heightened anonymity provided by the virtual environment [64]. In the case of social media platforms, such as Twitter, individuals have the option to adopt pseudonyms, helping to safeguard their real identity [65] and thereby helping to reduce these concerns. Further, during the pandemic, many people turned to social media as a way of combating depression and anxiety manifested from the event; discussing a range of different topics, including vaccines, along with individuals' perspectives, beliefs, and attitudes toward them [45, 66]. This makes social media a rich source of information on vaccine

hesitancy that can be curated and analysed, providing important insights that can be used to better understand this issue. Finally, from a more technical perspective, many social media platforms provide a dedicated Application Programming Interface (API), offering a flexible pathway for data retrieval that is often not available in conventional survey methodologies [67].

When it comes to vaccine hesitancy, several studies have explored this subject through the lens of social media. [68], for example, applied topic modelling to Twitter messages to gather information on the contributing factors of immunization uptake. That study identified various factors relating to access (e.g., location of vaccine), affordability (e.g., price of additional services), awareness (e.g., knowledge about vaccines), acceptance (e.g., perceived vaccine safety), activation (e.g., incentives), and assurance (e.g., protection) for variations in uptake. Similar applications of topic modelling include work by [35, 36, 69, 70]. A number of studies have also applied sentiment analysis to social media to gauge the general sentiment and attitudes of individuals towards vaccines and vaccination efforts [30, 71–74]. Work by [30, 31], for instance, used sentiment analysis to study the public's emotional stance surrounding the pandemic; with the public mainly having a negative view. More recently, research by [75] explored the association between vaccine hesitancy rates and socio-demographic characteristics derived from survey and social media data (i.e., Twitter). This study achieved notable accuracy levels for age (91%), gender (75%), and political ideology (77%). Moreover, a comparison of vaccine hesitancy figures from both survey data and Twitter posts across different time frames yielded Pearson's correlation coefficients in the range of 0.57 to 0.8.

Additional work by [26, 27] have studied misinformation campaigns about the pandemic. Those studies show that anti–vaccination communities on Twitter leaned mainly to the far right direction of the political spectrum, with references to websites and content with already questionable credibility. Moreover, [28, 29] studied the network interactions of members in vaccine–hesitant communities to understand their scale of impact, and the specific topics that were being propagated on social networks. Such studies, while informative, have mainly used the textual content embedded within social media posts to understand contributing factors towards hesitancy, missing the equally important social, demographic, and economic factors that also play a role in increased hesitancy rates [76].

In an attempt to address this issue, several studies have explored the use of such data for understanding the determinants of vaccine hesitancy. Studies by [77–79], for example, have identified low income, race, and level of education, to be important socio-demographic factors influencing hesitancy. Such work, however, have mainly relied on the use traditional surveys to collect this data; thus being exposed to some of their aforementioned issues with their use. Many countries collect large amounts of social, demographic, and economic data as part of national population census surveys, providing the ability to study such factors at scale. [80], for example, classified various socio-demographic variables into high and low socioeconomic groups to study topic prevalence within each group. That study showed that whereas the high group focused primarily on topics surrounding getting the vaccine, the low group mainly discussed an urgent need for medical and government support for the vaccine. Other work by [81] further built several regression models to understand the link between populations of unvaccinated persons and various socio–economic variables using census tracks data for the state of Texas in the US. That study reported main determinants to be neighborhoods with lower socio–economic standing and communities with signs of distrust in government.

Further work by [61] explored the performance of various machine learning algorithms using Twitter data for predicting vaccine hesitancy at the zip code level in the US. In that study, variables derived from Twitter messages (i.e., hashtags and sentiment score) were combined with different social, demographic, and economic variables (i.e., real estate value and

number of different health, educational, professional, scientific and technical service providing establishments) and used to predict vaccine hesitancy derived from Gallup poll survey data. That study found that while there was an improvement in model performance with the inclusion of Twitter data, overall performance was low, with reported root mean square error values between 0.3 and 0.4. More recent work by [82] further analysed the spatial and temporal impacts of neighbourhood variables on COVID-19 outbreaks. The results of that work showed the proportion of Hispanic residents, residents with earnings below the poverty line, and residents ages fifteen to twenty–four to have high correlation with high incidence of disease. Moreover, [51] extracted topics from both Twitter and survey data to compare co–variation in belief in vaccine hesitancy. Using tweets to infer stance (i.e., level of agreement or disagreement), the authors concluded that there was good qualitative agreement between the first principal component loading and scores using survey and Twitter data.

Most studies combining these forms of social data and social media data, while providing a more holistic view of vaccine hesitancy, do not directly examine the value of social media, as compared to the use of survey data, when trying to understand this issue. Several studies have examined this question for different areas, including, health [83], the economy [84], and entertainment [85], but to the authors' knowledge, there has been limited work with respect to vaccine hesitancy. The one exception is the study by [61]. However, as mentioned earlier, the reported accuracy values in that study were very low. Also, the data used were limited to real estate and types of establishments, providing an opportunity to explore other social, demographic, and economic variables, including those often investigated in related studies on vaccine hesitancy. Further, our study covers a much larger geographical area and integrates modern machine learning and deep learning approaches as part of our methodological workflow.

Methodology

Our research methodology involves the collection and preprocessing of two primary data sources: public opinion survey data concerning vaccine hesitancy and relevant Twitter data. These data were gathered within the geographical scope of the US, focusing specifically on understanding the attitudes and sentiments towards COVID-19 vaccination among US participants. The survey data captured the inclination of individuals receiving a COVID-19 vaccine. On the other hand, Twitter data, obtained from various repositories, were categorized into three distinct hesitancy stances: "pro" (favoring vaccines), "anti" (opposing vaccines), or "neutral" (neither favoring nor opposing vaccines). These two data sources were collected to facilitate a comparison between baseline models for different hesitancy groups. These models utilized only socio-demographic and economic variables, while others were augmented with either survey data or social media data. All data were collected at the county level, which aligns with the analytical focus of our study. The mixed methods matrix presented in Table 1 provides an overview of the steps undertaken, illustrating the integration of socio-demographic and economic data with survey data, and with Twitter data to enhance our understanding of vaccine hesitancy. These steps (i.e., data collection, data processing, and model development and comparison) are discussed in greater detail in the subsections that follow.

Data

As previously discussed, studies have reported multiple reasons influencing an individual's decision to not vaccinate. Therefore it was important to first identify these variables that contribute to vaccine hesitancy in order to develop our baseline models. To accomplish this, we first conduct an in-depth literature survey using online scholarly databases that included

Table 1. Mixed methods matrix showing the data, processing, and model development steps used in our study.

Data collection	Data processing	Model development and comparison	Outcome
COVID-19 vaccine hesitancy rates	(1) Extract rates for different hesitancy groups at the county level within each MSA study area.		(1) Hesitancy rates for different groups used to build our baseline models.
Socio-demographic and economic	(1) Review literature and identify factors related to vaccine hesitancy.	(1) Develop a set of baseline models, each representing a specific hesitancy rate group based on socio-demographic and economic variables.	(1) Baseline models used for explaining each hesitancy group.
	(2) Identify variables from relevant data sources.		
	(3) Extract variables at the county level within each MSA study area.		
	(4) Selection of most relevant variables for each hesitancy group.		
Survey	(1) Extract percentage of people within each hesitancy group at the county level within each MSA study area.	(1) Integrate survey data into each baseline model.	(1) Models highlighting the added value of incorporating survey data to each baseline model.
Social media	(1) Collect labelled tweets on vaccine hesitancy for different hesitancy stances.	(1) Use labelled tweets to build a multi-class model to classify tweets for each hesitancy stance.	(1) Models highlighting the added value of incorporating social media data to each baseline model.
	(2) Collect unlabelled tweets on vaccine hesitancy.	(2) Apply model to unlabelled tweets.	
	(3) Extract tweets at the county level within each MSA study area.	(3) Extract percentage of tweets within each hesitancy stance at the county level within each MSA study area.	
	(4) Integrate social media data into each baseline model.		

https://doi.org/10.1371/journal.pone.0301488.t001

Google Scholar [86], Web of Science [87], and Scopus [88] to identify variables relevant to vaccine hesitancy. Following the compilation of variables, applicable data sources were identified for the US as shown in Table 2. The collection of data was undertaken at the county scale. In instances where county-level data was unavailable, data at a higher spatial scale (e.g., zip code) was gathered and subsequently aggregated to the county level using a summative approach. With the exception of the social vulnerability variable [89], all other were data provided as counts of people. The social vulnerability variable is an index that measures the level of concern for a difficult roll-out on a range from 0 (lowest concern) to 1 (highest concern). This data includes multiple characteristics of the people that live in counties.

County level vaccine hesitancy rates were collected from the US Department of Health and Human Services (HHS) [97]. This data consisted of estimated COVID-19 hesitancy rates for each county in the US. To generate these estimates, the data initially utilized the Census Bureau's Household Pulse Survey (HPS) data at the state level and subsequently extrapolated county-level rates using the Census Bureau's 2019 ACS Public Use Microdata Sample (PUMS). HPS participants were asked if they would receive the COVID-19 vaccine when it became available. Five responses were captured: "definitely get a vaccine", "probably get a vaccine", "unsure", "probably not get a vaccine", and "definitely not get a vaccine". Responses were used by HHS to compute data for three hesitancy groups: "strongly hesitant," "hesitant," and "hesitant or unsure." The *strongly hesitant* group refer to people that stated they would "definitely not" receive a COVID-19 vaccine. The *hesitant* group refer to people that indicated that they would "probably not" or "definitely not" receive a COVID-19 vaccine. Finally, the *hesitant or unsure* group refer to people that stated they would "probably not" or "unsure" or "definitely not" receive a COVID-19 vaccine. Further, we use the HHS data to derive two additional group measures. The *unsure* group was derived by subtracting the percentage of people that

Table 2. Data sources used in our study.

Variable	Description	Data source	Reference	Year(s)
Age	Number of persons in different age groups	American Community Survey, US Census Bureau	[90]	2019
Ethnicity	Number of persons in different ethnic groups	American Community Survey, US Census Bureau	[90]	2019
Education	Number of persons at different education levels	American Community Survey, US Census Bureau	[90]	2019
Cohabitation	Number of couples cohabiting	American Community Survey, US Census Bureau	[90]	2019
Occupation	Number of people in different job classifications	US Bureau of Labor Statistics	[91]	2021
Employment	Number of people employed and unemployed	US Bureau of Economic Analysis	[92]	2021
Income	Number of people in different income groups	American Community Survey, US Census Bureau	[90]	2019
Political party	Number of people that have voted for a specific political group	Harvard Dataverse	[93]	2020
COVID-19 cases	Number of COVID-19 cases	The New York Times	[94]	2021
COVID-19 vaccination	Percentage of people that have received the primary dose of a COVID-19 vaccine	Center for Disease Control	[95]	2021
COVID-19 vaccination	Percentage of people that have received the complete series of a COVID-19 vaccine	Center for Disease Control	[95]	2021
Social vulnerability	Social Vulnerability Index	Center for Disease Control	[96]	2021
Social vulnerability	Level of concern of vaccine rollout (value between 0 and 1)	Surgo Ventures	[89]	2021
Survey	Percentage of people that are vaccine hesitant	US Department of Health and Human Services	[97]	2021
Survey	Percentage of people that answered yes to specific questions on vaccines	Delphi survey, Carnegie Mellon University	[98]	2022
Social media	Twitter posts labelled as Pro-vaccine, Antivaccine, and Neutral	Twitter	[99]	2015 to 2021
Social media	Twitter posts labelled as Pro-vaccine, Antivaccine, and Neutral	Twitter	[100]	2015 to 2020
Social media	Twitter posts labelled as Pro-vaccine, Antivaccine, and Neutral	Twitter	[101]	2019 to 2021
Social media	Twitter posts labelled as Pro or Neutral	Twitter	[102]	2017 to 2020
Social media	Unlabelled Twitter posts	Twitter	[103]	2016 to 2021
Boundaries	Geographic boundaries—217 counties accross 10 metropolitan statistical areas	American Community Survey, US Census Bureau	[90]	2021

https://doi.org/10.1371/journal.pone.0301488.t002

answered that they would "probably not" or "definitely not" receive a COVID-19 vaccine from the *hesitant or unsure* group. The *probably not* group was derived by subtracting the percentage of people that stated they were "unsure" or "definitely not" receive a COVID-19 vaccine from the *hesitant or unsure* group. Computed values for these hesitancy groups represent our five target variables in this work.

Survey public opinion data were collected from the Delphi Survey, a product from the Delphi group at Carnegie Melon University [98]. The survey asked US participants, if a vaccine to prevent COVID-19 were offered to you today, would you choose to get vaccinated? Responses were: (1) Yes I would definitely choose to get vaccinated, (2) Yes I would probably choose to get vaccinated, (3) No I would probably not choose to get vaccinated, or (4) No I would definitely not choose to get vaccinated. This data were used by the Delphi group to compute two measures of hesitancy, and made available for public use. The "somewhat' group is computed as the proportion of respondents who answered "no" or "I don't know" to the second question

divided by the total number of respondents who gave any answer to the first question. The "all" group is computed as the proportion is the number of respondents who answered "yes", "no" or "I don't know" to the second question divided by the total number of respondents who gave any answer to the first question. Percentages of the population in each group were provided for each US county. There represent our two survey variables.

Our work also leverages two types of Twitter data. Labeled tweets were collected from several sources [99–102], which each categorized them as one of three different hesitancy stances:"pro", "anti", or "neutral". The count distribution of these tweets were 8,213, 2,322, 7,017 for "pro", "anti", and "neutral" stances respectfully. In some instances, only tweet IDs were provided from the online repositories, whereas the corresponding tweeted message was required. To collect these messages, the tweet ID was utilized to access the Twitter API, and the corresponding text message was saved.

Unlabelled Twitter data consisted of tweets that include not only vaccine hesitancy-related tweets related to COVID-19, but also discussions about other diseases such as measles, mumps, and rubella, which similarly sparked online conversations on social media during this time frame [99, 104, 105]. This diversity in hesitancy across multiple diseases is also reflected in the labelled data as well. The unlabelled data were collected from the Twitter platform using their API. A set of keywords that included "vaccine", "vax", "vaccine hesitancy", "vaccine hesitant", "anti-vax", "anti-vaxx", "antivax", and "antivaxx", were employed to query the API. These keywords were chosen based on previous research that delved into vaccine hesitancy through social media analysis. For instance, [15] utilized both Twitter and Facebook data and concluded that vaccine-hesitant individuals are more likely to post vaccine misinformation online compared to other groups. Similarly, [106] reported a significant presence of vaccine-hesitant groups across popular social media platforms, posting anti-vaccination messages. Only English tweets within the US were considered in this work. All personally identifiable information were removed from both labelled and unlabelled tweets and records were anonymized to protect the privacy of individuals. Additionally, to ensure data integrity and consistency, a rigorous data preprocessing pipeline was implemented, including steps such as removing duplicates, filtering out irrelevant tweets, and standardizing text formats. Moreover, procedures were carefully taken to ensure compliance with Twitter's terms of service and use of data usage agreements.

Finally, we collect administrative boundary data for the ten of the most populous Metropolitan Statistical Areas (MSAs) in the US [90]. It is important to note that these MSAs serve as our primary study areas. They represent well-defined geographic regions with substantial economic and social significance, each with their own specific local behaviour and regional trend [107]. Further, recognizing that social media use tends to be concentrated in regions with larger populations and greater technological access, known as the digital divide [108–110], the use of MSA as our study areas helps to reduce such concerns. The MSAs are Atlanta-Sandy Springs-Alpharetta (GA), Chicago-Naperville-Elgin (IL-IN-WI), Dallas-Fort Worth-Arlington (TX), Washington-Arlington-Alexandria (DC-VA-MD-WV), Houston-The Woodlands-Sugar Land (TX), Los Angeles-Long Beach-Anaheim (CA), Miami-Fort Lauderdale-Pompano Beach (FL), New York-Newark-Jersey City (NY-NJ-PA), Philadelphia-Camden-Wilmington (PA-NJ-DE-MD), and Phoenix-Mesa-Chandler (AZ). These will henceforth be labelled as Atlanta (Atl), Chicago (Chi), Dallas (Dal), Washington DC (DC), Houston (Hou), Los Angeles (LA), Miami (Mia), New York (NYC), Philadelphia (Phl), and Phoenix (Phx), respectfully. Readers interested in more details on the specific collection and processing steps for the various data used in this research are referred to their specific reference in Table 2.

Data processing

Following the extraction of socio-demograhic and economic variables (discussed in the previous section), with the exception of the the social vulnerability variable, all other variables were transformed from the absolute number of individuals within that variable group to a percentage value. This was done using total population data for counties provided by the American Community Survey [90]. Since no information is provided on the number of people or percentage of the population within each level of concern sub-category for social vulnerability, this data could not be transformed into a percentage value. Nonetheless, many studies have used this variable as a seminal measure to understand vaccine hesitancy (e.g., [111, 112]); thus, the decision was made to keep it as a variable in this work.

After variables were transformed, there was need to identify the most relevant subset of variables for each baseline hesitancy target group (i.e., strongly hesitant, hesitant, hesitant and unsure, unsure, and probably not). As emphasized by [113], this process is important for eliminating irrelevant, noisy, or unreliable variables, ultimately improving predictions and/or minimizing model complexity. Moreover, this approach mitigates the risk of over-fitting and enhances model runtime efficiency. For determining the optimal subset of explanatory variables for modeling, the BorutaSHAP [114] algorithm was used. This process yielded five distinct subsets of variables.

The BorutaSHAP algorithm combines the advantages of both the Boruta [115] and SHapley Additive exPlanations (SHAP) [116] algorithms to identify the most optimal subset of explanatory variables. The Boruta algorithm operates through iterative comparison of the importance of original variables against shadow variables, which are created by shuffling the original variables. Variables demonstrating significantly lower importance than their shadow counterparts are excluded from the variable set, while those performing notably better than the shadow variables are retained [115]. In the context of BorutaSHAP, the SHAP metric is utilized to ascertain variable relevance [117], often resulting in improved overall accuracy compared to Boruta [118]. Prior research has indicated that BorutaSHAP serves as a reliable feature selection technique [119–122], making it suitable for application within this study.

Concerning the Twitter data, all unlabelled data were spatially clipped to each respective MSA study area resulting in an approximate count of five million tweets. To ensure data quality, tweets further underwent a cleaning process involving the removal of URLs, emails, and usernames, the expansion of contracted words, and the replacement of emojis with their corresponding textual descriptions as suggested by previous related work [123–125].

The Delphi survey data consisted of two hesitancy groups: 'somewhat' and 'all'. The percentage of each group per county was used to compute the corresponding percentage for the respective MSA. This aggregation involved utilizing the weighted sum of the population percentages within each MSA, and computation for each MSA as follows. Let $C_i \in \mathcal{C}\ \forall i \in 1, \ldots, n$ where $\mathcal{C}$ is the set of all counties in an MSA. The percentage of interest within the metropolitan statistical area, $H(M)$, is given by:

$$H(M) = \frac{\Sigma_{i=1}^{n}(H(C_i) \times P(C_i))}{\Sigma_{i=1}^{n} P(C_i)},$$

where $H(M)$ represents the percentage of the population within an MSA, $H(C_i)$ is the percentage of interest for the i^{th} county of the MSA, and $P(C_i)$ is the total number of people for the i^{th} county of the MSA. Similar to the Twitter data, the computed percentages of hesitant population at the MSA level were added as an attribute to each county within the respective MSA.

Model development and comparison

Leveraging the Twitter labeled dataset with text, a classification model was constructed utilizing BERTweet [126], a pre-trained English tweet language model [127]. The model development followed an 80/20 training/testing split strategy. For model refinement and the identification of optimal hyperparameters, the Ray-Tune optimization framework, employing a population-based scheduler, was used in Python [128]. The resulting model achieved an F-score of 0.83, a metric value consistent with previous studies indicating a good model fit [38, 101, 129, 130]. The developed classification model was then applied to unlabeled data, classifying tweets as "pro", "anti", and "neutral" respectfully. Following this, the percentage of tweets per MSA, for each hesitancy group, was computed. These values were then added as an attribute to all counties within the respective MSA. This was done in order to account for distributional differences in tweets within MSAs, and to identify their broad regional hesitancy patterns. Moreover, aggregating the data to a higher geographic scale allows for a larger sample size, improving the statistical power of the analysis. This is crucial for making confident inferences and identifying meaningful correlations with other variables used in this study.

Next, a series of baseline models, that only contain the social, demographic, and economic explanatory variables were developed and compared with models that were augmented with either the survey or social media data. Three distinct modeling techniques were employed for this comparison: linear regression, random forest regression, and XGBoost regression. These methods were specifically chosen as they allow for the comparison of similar performance metrics, namely, R^2 and mean absolute value. Additionally, these modeling techniques are widely utilized in similar contexts and are all available within various open-source Python packages (i.e., Scikit-Learn [131] for linear regression and random forest, and xgboost [132] for XGBoost), which were utilized in this study.

For each modeling method, a set of three models aimed at predicting hesitancy were built, one for each target variable (i.e., 'strongly hesitant,' 'hesitant,' 'hesitant and unsure,' 'unsure,' and 'probably not'). To illustrate, for the 'strongly hesitant' group of models, there were three baseline models, each corresponding to the three modeling techniques employed. Additionally, there were three models that incorporate our survey variables and another three that incorporate our social media variables for the same target variable, again aligning with the three modeling techniques. To assess model performance, each model underwent 500 randomized 80/20 training/testing data splits, and their average adjusted R^2 accuracy was recorded.

Similar to recent work by [61], a significance analysis was also carried out to determine the effectiveness of the best model. We used the Mann-Whitney U test statistic [133] and compared the baseline models against their social media and survey counterparts to determine whether the performance of these later models can be attributed to chance alone. Specifically we compute the p-value from this test statistic using the distribution of average R^2 values. To address the potential for inflated significance due to multiple testing, we applied the Bonferroni correction on the p-value threshold of 0.05. This adjustment reduces the p-value threshold below the 0.05 threshold to account for multiple tests. We found even with the correction that that there was no change in the significance of our results before and after applying the Bonferroni correction in terms of comparing adjusted R^2 value of the models.

Moreover, we assess the extent to which models can be applied generally. To achieve this, we follow a similar approach to [134], initially partitioning the data by MSA. Subsequently, we employed all three models (baseline, social media, and survey) for every technique and each hesitancy group within each MSA study area. This allowed for comprehensive comparisons

across regions. The evaluation of each model's effectiveness was based on their respective R^2 values.

Results

The primary objective of this study was to evaluate the utility of social media data as a proxy measure for understanding vaccine hesitancy. This objective is assessed in the first two subsections that follow, where we compare the performance of models utilizing social media data with those using survey data. Through this analysis, we aim to determine the effectiveness of social media data in capturing and predicting vaccine hesitancy trends, along with exploring the influence of geographic variation.

The secondary goal was to explore the determinants of vaccine hesitancy using modern machine learning approaches. This objective is assessed in the final subsection, where we examine the factors contributing to vaccine hesitancy identified by the developed models. By leveraging machine learning techniques, we aim to identify key determinants such as demographic variables and socioeconomic factors that influence individuals' attitudes towards vaccination.

Model performance and significance analysis

The information presented in Table 3 provides an overview of the performance metrics for various developed models and techniques. These metrics encompass average root mean squared error (RMSE), adjusted R^2, and the percentage change in adjusted R^2 with the addition of the social media and survey data into the baseline models. The RMSE values demonstrate strong model fit across all hesitancy models, particularly for the unsure, probably not, and strongly hesitant models. However, a closer examination of models' adjusted R^2 values reveal larger variability among these values. Specifically, the range of adjusted R^2 values spans from 50% to 95%, highlighting differences both between models and across techniques. The linear approach has the largest performance range, with adjusted R^2 values spanning from 50.4% to 89.5%. The adjusted R^2 range for both random forest and XGBoost is comparable, lying between 75.8% and 94.1% for random forest, and 73.1% and 92.5% for XGBoost. Notably, XGBoost is the best performing method with an average of 87.2% and a standard deviation of 0.06. Nonetheless, these figures align closely with those of random forest, having an average of 86.9% and a standard deviation of 0.07. Furthermore, XGBoost consistently outperforms the other two methods across all models, while random forest also demonstrates superior performance compared to the linear method for all models.

Turning to the utility of models utilizing social media data versus those leveraging survey data, Table 3 highlights that, except for the probably not model, models utilizing survey data consistently outperform both baseline models and those utilizing social media data across all approaches. For the probably not model, the performance disparity between the XGBoost model using social media data and its survey data counterpart is marginal at 0.16% higher for the former. Likewise, with the exception of the unsure model, models using social media data outperform all baseline models. Concerning the unsure model, linear and random forest baseline models exhibit better performance, achieving slight increases in R^2 values of 5.53% and 0.02%, respectively, in comparison to the social media model. Notably, the performance values for the linear method remain relatively low, ranging from 55% to 57%. A comparison between the performance of social media models and those utilizing survey data underscores differences in the range of 0.16% to 12%. The most substantial differences are observed for the linear method, with disparities for random forest and XGBoost ranging from 0.15% to 3.1%. More generally, our findings show that with the additional of social media or survey data, model

Table 3. Model performance (Bolded values represent the best performing model for each approach).

Method	Model	RMSE	Adjusted R^2	% Improvement in Adjusted R^2
Hesitant				
Linear	Baseline	0.0216	0.7384	N/A
Linear	Social media	0.0163	0.8480	10.97
Linear	Survey	0.0135	**0.8952**	15.68
Random Forest	Baseline	0.0146	0.8771	N/A
Random Forest	Social media	0.0123	0.9132	3.62
Random Forest	Survey	0.0103	**0.9392**	6.21
XGBoost	Baseline	0.0162	0.8491	N/A
XGBoost	Social media	0.0125	0.9113	6.23
XGBoost	Survey	0.0115	**0.9249**	7.58
Unsure				
Linear	Baseline	0.0091	0.5595	N/A
Linear	Social media	0.0097	0.5042	-5.53
Linear	Survey	0.0090	**0.5673**	0.79
Random Forest	Baseline	0.0069	0.7582	N/A
Random Forest	Social media	0.0068	0.7580	-0.02
Random Forest	Survey	0.0068	**0.7595**	0.13
XGBoost	Baseline	0.0071	0.7311	N/A
XGBoost	Social media	0.0070	0.7393	0.82
XGBoost	Survey	0.0069	**0.7503**	1.92
Hesitant or unsure				
Linear	Baseline	0.0247	0.7616	N/A
Linear	Social media	0.0240	0.7704	0.88
Linear	Survey	0.0172	**0.8840**	12.23
Random Forest	Baseline	0.0184	0.8673	N/A
Random Forest	Social media	0.0168	0.8899	2.26
Random Forest	Survey	0.0146	**0.9145**	4.72
XGBoost	Baseline	0.0204	0.8393	N/A
XGBoost	Social media	0.0172	0.8835	4.42
XGBoost	Survey	0.0148	**0.9129**	7.35
Probably not				
Linear	Baseline	0.0090	0.6741	N/A
Linear	Social media	0.0075	0.7698	9.58
Linear	Survey	0.0062	**0.8412**	16.71
Random Forest	Baseline	0.0051	0.8942	N/A
Random Forest	Social media	0.0039	0.9391	4.49
Random Forest	Survey	0.0038	**0.9416**	4.74
XGBoost	Baseline	0.0061	0.8488	NA
XGBoost	Social media	0.0046	**0.9164**	6.75
XGBoost	Survey	0.0046	0.9147	6.59
Strongly hesitant				
Linear	Baseline	0.0141	0.7270	N/A
Linear	Social media	0.0120	0.8008	7.38
Linear	Survey	0.0095	**0.8734**	14.64
Random Forest	Baseline	0.0112	0.8308	N/A
Random Forest	Social media	0.0097	0.8744	4.36
Random Forest	Survey	0.0084	**0.9051**	7.43

(*Continued*)

Table 3. (Continued)

Method	Model	RMSE	Adjusted R^2	% Improvement in Adjusted R^2
XGBoost	Baseline	0.0121	0.8023	N/A
XGBoost	Social media	0.0102	0.8597	5.74
XGBoost	Survey	0.0091	**0.8864**	8.41

https://doi.org/10.1371/journal.pone.0301488.t003

performance can be improved by as much as 17%, dependent on the specific model and technique being used.

Additionally, Table 4 shows the outcomes of the significance analysis. This table demonstrates that models developed for unsure, employing both social media and survey data, exhibit

Table 4. Significance analysis.

Method	Model	P-value
Hesitant		
Linear	Social media	9.57×10^{-108}
Linear	Survey	8.43×10^{-155}
Random Forest	Social media	4.47×10^{-18}
Random Forest	Survey	2.05×10^{-88}
XGBoost	Social media	3.57×10^{-59}
XGBoost	Survey	5.49×10^{-99}
Unsure		
Linear	Social media	1
Linear	Survey	9.99×10^{-1}
Random Forest	Social media	9.99×10^{-1}
Random Forest	Survey	9.99×10^{-1}
XGBoost	Social media	9.99×10^{-1}
XGBoost	Survey	3.78×10^{-1}
Hesitant or unsure		
Linear	Social media	3.07×10^{-24}
Linear	Survey	1.79×10^{-144}
Random Forest	Social media	1.56×10^{-4}
Random Forest	Survey	1.42×10^{-50}
XGBoost	Social media	1.90×10^{-25}
XGBoost	Survey	1.00×10^{-107}
Probably not		
Linear	Social media	2.33×10^{-78}
Linear	Survey	3.26×10^{-128}
Random Forest	Social media	1.47×10^{-43}
Random Forest	Survey	2.98×10^{-52}
XGBoost	Social media	3.42×10^{-58}
XGBoost	Survey	2.79×10^{-58}
Strongly hesitant		
Linear	Social media	1.72×10^{-85}
Linear	Survey	5.43×10^{-139}
Random Forest	Social media	2.68×10^{-13}
Random Forest	Survey	8.84×10^{-62}
XGBoost	Social media	9.05×10^{-28}
XGBoost	Survey	4.06×10^{-80}

https://doi.org/10.1371/journal.pone.0301488.t004

p-values well below the 0.05 significance threshold. Compared with the performance outcomes shown in Table 3, it is evident that the improvements in performance in these unsure models are minimal compared to other hesitancy models, ranging from 0.13% to 1.92%. Consequently, we accept the null hypothesis in this instance, concluding that these models' results can be attributed to random chance. Conversely, all other models present significantly small p-values, indicating robustness and results that are not contingent on random chance. These findings parallel the performance metrics in Table 3, wherein the models with the weakest performance are predicated on linear methods. In the context of XGBoost, with the exception of the probably not hesitancy model, this method continues to outperform all others.

Model generalizability to metropolitan statistical areas

Table 5 shows the performance, measured by R^2, of the models applied to their respective MSAs. The table highlights the performance variation across each MSA. Negative R^2 values in this context indicate that the model's performance is below average, reflecting poor performance. For Miami, Los Angeles, and Phoenix, the performance of all models are particularly poor in this respect. These MSAs had 4, 2, and 2 counties respectively compared to other the number of counties in other MSAs that were in the range of 12 to 54. In this case, the lower amount of counties participating in the training data may skewed the performance to MSAs with a larger number of counties. For the remaining MSAs, models' performance is much higher, with values exceeding 82% on average for random forest and XGBoost. This trend aligns with the patterns discussed in the previous section, where XGBoost consistently outperformed linear regression and random forest methods. It is important to highlight that linear regression is consistently the least effective modeling method in terms of performance.

Notably, while the distinction between the performances of XGBoost and random forest is evident, the margin of differentiation is comparatively small. Furthermore, in line with earlier findings, models incorporating survey data consistently demonstrate superior performance when juxtaposed against baseline models or models relying on social media data.

Determinants of vaccine hesitancy

Vaccine hesitancy poses a complex and multifaceted challenge that is influenced by multiple contributing factors. Those factors pertaining to the models investigated within this study are outlined in Table 6. The table highlights the diverse array of factors operating within distinct models. In a descending order based on the number of variables are the hesitant or unsure (16 variables), unsure (14 variables), hesitant (13 variables), probably not (7 variables), and strongly hesitant (7 variables) models respectfully. The factors associated with the hesitant model primarily revolve around levels of education, cohabitation, occupation, income, political inclinations, as well as the prevalence of COVID-19 cases and vaccination rates. A similar pattern is observed in the unsure model, which also includes ethnicity. On the other hand, the hesitant or unsure model encompasses a broader spectrum of factors, utilizing most variables examined in this study, and span categories that include age, ethnicity, education, cohabitation, occupation, income, political leaning, vaccine distribution, and social vulnerability. Further, the probably not and strongly hesitant models are very similar, sharing factors linked to education, occupation, and income. The probably not model also includes employment status as an important factor.

Discussion

Vaccine hesitancy is a worldwide phenomena that poses a significant challenge to public health efforts to control or eradicate preventable, but potentially harmful diseases [135]. While not a

Table 5. MSA model performance (Bolded adjusted R^2 values represent the best performing model for each modelling technique and MSA).

Method	Model	Atl	Chi	Dal	DC	Hou	LA	Mia	NYC	Phl	Phx
Hesitant											
Linear	Baseline	-0.331	0.283	0.211	-1.973	0.495	-27.81	-1.855	0.449	0.273	**0.941**
Linear	Social media	0.517	**0.658**	0.530	0.076	0.650	**0.591**	-45.77	0.699	**0.480**	-0.259
Linear	Survey	**0.613**	0.633	**0.631**	**0.309**	**0.755**	-8.162	-23.16	**0.782**	0.376	0.178
Random Forest	Baseline	**0.889**	0.851	0.840	0.535	0.709	-19.49	-1.499	0.945	0.787	**0.537**
Random Forest	Social media	0.877	0.885	**0.940**	0.723	0.839	-0.223	**0.684**	**0.961**	**0.963**	-4.634
Random Forest	Survey	0.806	**0.971**	0.927	**0.932**	**0.924**	**0.621**	-5.010	0.958	0.954	-1.668
XGBoost	Baseline	**0.918**	0.952	0.763	**0.962**	0.659	0.931	**0.988**	0.970	0.985	**0.999**
XGBoost	Social media	0.914	**0.995**	0.934	0.899	**0.901**	0.974	-1.738	0.957	**0.992**	-3.026
XGBoost	Survey	0.871	0.921	**0.979**	0.849	0.762	**0.996**	0.570	**0.972**	0.987	**0.999**
Unsure											
Linear	Baseline	0.515	0.342	0.108	0.554	0.237	-11.75	-0.298	0.512	**0.451**	-25.29
Linear	Social media	0.468	0.424	0.062	**0.581**	0.301	**0.536**	-1.613	**0.526**	0.433	-7.787
Linear	Survey	**0.644**	**0.535**	**0.222**	0.479	**0.422**	-8.372	**0.348**	0.464	0.450	-14.36
Random Forest	Baseline	0.808	0.732	0.795	0.939	**0.941**	-2.660	0.953	0.896	0.846	-1.314
Random Forest	Social media	0.787	0.835	**0.954**	**0.949**	0.863	-2.022	0.952	**0.948**	**0.935**	-6.930
Random Forest	Survey	**0.848**	**0.896**	0.949	0.846	0.719	-16.27	**0.976**	0.904	0.929	-6.406
XGBoost	Baseline	**0.875**	0.941	0.782	0.889	0.955	-4.445	0.981	0.906	0.961	0.988
XGBoost	Social media	0.812	0.447	**0.991**	**0.959**	**0.993**	**0.940**	**1.000**	0.911	0.961	-1.687
XGBoost	Survey	0.809	**0.970**	0.976	**0.959**	0.807	0.885	0.989	**0.980**	**0.982**	**0.997**
Hesitant or Unsure											
Linear	Baseline	-0.054	0.250	-0.226	-0.005	0.541	-12.38	-2.045	0.413	-0.128	-6.389
Linear	Social media	0.542	0.569	0.343	0.437	**0.895**	**0.699**	-16.80	0.261	**0.007**	-28.28
Linear	Survey	**0.700**	**0.654**	**0.465**	**0.604**	0.773	-3.452	-1.801	**0.608**	-0.557	-0.669
Random Forest	Baseline	0.839	**0.936**	0.910	0.672	**0.953**	-4.641	-0.980	**0.900**	0.768	-1.139
Random Forest	Social media	0.854	0.904	0.871	0.875	0.935	-4.434	**0.821**	0.858	**0.836**	-0.436
Random Forest	Survey	**0.909**	0.848	**0.911**	**0.889**	0.939	**0.358**	0.494	0.868	0.695	**0.269**
XGBoost	Baseline	0.932	0.970	**0.970**	0.964	0.831	**0.969**	0.987	0.958	**0.920**	0.907
XGBoost	Social media	0.841	**0.999**	0.913	0.963	0.713	0.152	**0.995**	0.899	0.819	-14.84
XGBoost	Survey	**0.952**	0.901	0.920	**0.978**	**0.974**	0.022	-0.324	**0.990**	0.905	**0.998**
Probably not											
Linear	Baseline	-1.001	0.253	-0.171	-1.247	0.315	-0.938	-1.808	0.413	0.607	-167.1
Linear	Social media	0.305	0.327	0.065	**0.275**	0.452	-7.702	-79.68	0.492	**0.710**	-30.51
Linear	Survey	**0.373**	**0.727**	**0.388**	0.252	**0.843**	-13.82	-19.21	**0.741**	0.339	-136.2
Random Forest	Baseline	0.756	0.856	0.879	0.650	0.882	**0.770**	-0.479	0.972	0.927	-1.385
Random Forest	Social media	**0.895**	0.895	0.871	0.875	**0.935**	0.437	**0.731**	**0.979**	0.877	**0.156**
Random Forest	Survey	0.807	**0.970**	**0.885**	**0.959**	0.878	-1.575	0.590	0.980	**0.986**	-6.740
XGBoost	Baseline	0.689	0.941	**0.978**	0.609	0.955	**0.914**	**0.915**	**0.982**	0.928	**0.259**
XGBoost	Social media	0.922	0.447	0.767	**0.977**	**0.993**	0.754	0.732	0.968	**0.981**	-4.122
XGBoost	Survey	**0.952**	**0.970**	0.976	0.931	0.807	0.781	-0.492	0.973	0.922	-3.368
Strongly Hesitant											
Linear	Baseline	-0.134	0.511	0.042	-1.457	0.529	-398.6	-2.061	0.726	0.382	-3.219
Linear	Social media	0.572	0.833	0.501	0.030	**0.676**	-12.52	-30.01	0.710	**0.324**	-9.771
Linear	Survey	**0.578**	**0.907**	**0.532**	**0.088**	0.654	-116.7	-12.20	**0.802**	0.215	-2.399
Random Forest	Baseline	0.809	0.908	0.794	0.322	0.943	-174.3	-11.65	0.934	0.783	-2.019
Random Forest	Social media	**0.845**	**0.969**	0.594	0.880	**0.906**	-23.54	**0.478**	**0.942**	**0.904**	-0.305
Random Forest	Survey	0.531	0.822	**0.896**	**0.937**	0.866	-11.39	-0.404	0.939	0.697	**0.367**

(*Continued*)

Table 5. (Continued)

Method	Model	Atl	Chi	Dal	DC	Hou	LA	Mia	NYC	Phl	Phx
XGBoost	Baseline	**0.875**	0.907	0.975	0.851	0.659	-324.4	**0.972**	0.842	0.992	0.987
XGBoost	Social media	0.812	0.736	**0.994**	-0.325	**0.901**	0.471	-0.233	0.707	**0.993**	**0.999**
XGBoost	Survey	0.809	**0.996**	0.867	**0.917**	0.762	**0.811**	0.642	**0.872**	0.901	-4.983

https://doi.org/10.1371/journal.pone.0301488.t005

new issue, the problem has become more endemic in wake of the COVID-19 pandemic, leading to millions of people not being vaccinated globally. [136] estimates that within the first year of the pandemic alone, almost 20 million deaths were averted due to vaccines. Similarly, within the US, between 2020 and 2021, vaccines were estimated to prevent approximately 27

Table 6. Factors related to vaccine hesitancy.

Variable group	Variable	Hesitant	Unsure	Hesitant or unsure	Probably not	Strongly hesitant
Age	Age less than 18 years			✓		
Ethnicity	Black population		✓			
Ethnicity	Asian population			✓		
Education	High school education			✓		
Education	Some college degree	✓		✓	✓	✓
Education	Associate degree					✓
Education	Bachelor degree		✓			
Cohabitation	Couple cohabitation			✓		
Occupation	Manufacturing	✓		✓	✓	✓
Occupation	Professional, scientific, and management, and administrative and waste management service	✓	✓	✓	✓	✓
Occupation	Wholesale trade	✓		✓	✓	✓
Occupation	Sales and office	✓				✓
Occupation	Self employed		✓			
Occupation	Finance and insurance, and real estate and rental and leasing		✓			
Occupation	Transportation and warehousing, and utilities			✓		
Occupation	Private wage and salary workers				✓	✓
Employment	People employed				✓	
Income	Per capita income	✓	✓	✓	✓	✓
Income	Income between 25k and 50k	✓		✓		
Political inclination	Democrat	✓	✓	✓		
Political inclination	Green					
Political inclination	Libertarian	✓	✓	✓		
Political inclination	Other political party	✓	✓	✓		
COVID-19 cases	COVID-19 cases	✓				
COVID-19 vaccination	Completed primary COVID-19 vaccination series	✓	✓			
COVID-19 vaccination	Completed vaccine series against COVID-19		✓	✓		
COVID-19 vaccination	Age 65 years and over that have been administered the first vaccine dose		✓			
Social vulnerability	Social Vulnerability Index		✓			
Social vulnerability	Level of concern	✓	✓	✓		

https://doi.org/10.1371/journal.pone.0301488.t006

million infections, 1.6 million hospitalizations, and 235,000 deaths [137], with more than 300,000 deaths being prevented the following year [138]. Further, [139] estimates that for every one percent decrease in vaccine hesitancy, as much as 45 deaths per million people could have been prevented during the pandemic, making understanding vaccine hesitancy of key importance for human survival.

One key issue with understanding vaccine hesitancy is with the speed at which data can be gathered and analyzed to provide key insights as events unfold. Traditionally, surveys have been used for this purpose. However, as as discussed earlier on in this paper, this data comes with it's own set of caveats. In this study we assessed the use of social media data as a potential source of insights on vaccine hesitancy, helping to improve the performance of models used for understanding the determinants surrounding the reluctance to vaccinate when compared to the use of survey data. However, it is important to note that there are also various limitations with the use of social media data on it's own to understand this issue. For example, certain demographics or socioeconomic groups may be overrepresented or underrepresented [140]. Related is the issue of access bias; not all segments of the population have equal access to social media platforms or may not actively engage in online discussions. This can result in underrepresentation of certain demographics, such as older adults or individuals from low-income communities, in social media data [141]. Social media algorithms may also prioritize content that aligns with some users existing beliefs and preferences, reinforcing pre-existing biases and limiting exposure to diverse perspectives [35]. Moreover, social media data analysis is often conducted in specific languages, which may introduce language bias. Insights drawn from social media data may not be applicable to populations that primarily communicate in different languages, limiting the generalizability of findings [142]. Nevertheless, although various data sources and methodologies may present slightly different viewpoints, together they contribute to a thorough comprehension of the extent and reasons behind vaccine hesitancy at a population level. These varied perspectives function as integral components of a broader puzzle, facilitating the synthesis of insights necessary for developing effective strategies to tackle vaccine hesitancy.

Our results demonstrate that the addition of survey data consistently provides improved model performance compared to social media data across various forms of hesitancy (i.e., hesitant, unsure, probably not, strongly hesitant) and approaches (i.e., linear regression, random forest, XGBoost). However, it's noteworthy that in some cases, this improvement was marginal, particularly when the XGBoost and random forest techniques were used. This is evidenced by our significance analysis, which demonstrate the robustness of these models in capturing the complexities of hesitancy dynamics.

Additionally, while the generalizability of models to metropolitan statistical areas (MSAs) was generally satisfactory, there were instances of poor MSA performance, with variations observed across different methods. XGBoost still continued to be a robust performer relative to other methods, especially when used in conjunction with survey data. This reinforces the importance of survey data in improving model accuracy beyond the limitations of baseline or social media-derived models. Further, within the same locales and for the top performing models, similar to before, there was marginal difference in model performance between models that use survey data and those that use social media data, highlighting the promise of social media data as a valid source data for understanding vaccine hesitancy. Moreover, the observed performance variations across different MSA locales indicate the influence of additional factors and unique attributes associated with each locale on model outcomes.

Further, congruent with other work, we found age, ethnicity, education, cohabilitation, occupation, employment, income, political leaning, number of disease cases, and the distribution and level of concern for vaccine distribution to be key factors in understanding vaccine

hesitancy. It's important to note, however, that these factors were operationalized differently across the various hesitancy models in this study. This underscores the necessity to consider these nuances when interpreting and applying the results to address vaccine hesitancy effectively. Age, in particular, was only applicable to the hesitant or unsure model, and specifically, people less than 18 year of age. Previous work have explored the relationship between age and ethnicity. For instance, a study by [143] in Ireland and the UK found heightened reluctance to vaccine in age groups 35-45 and 18-24 for Ireland and UK respectfully. [144] investigated schoolchildren aged 9-18 and found considerable indecision (37%) about vaccination, while 12.9% answered that they would opt-out to getting a COVID-19 vaccine. The main reasons given for their reluctance to become vaccinated included distrust of vaccines, government agencies promoting their uptake, apprehensions about side effects, and perceptions of low personal risk. Notably, those opting against vaccination demonstrated higher degrees of marginalization and skepticism towards vaccine information, highlighting the necessity for greater government intervention in addressing these concerns.

The role of ethnicity, specifically the representation of black and Asian populations (primarily originating from East and Southeast Asia, as well as the Indian subcontinent), emerged as a significant determinant within the unsure and hesitant or unsure models. Previous research has extensively explored ethnicity, particularly within the context of BAME (black, Asian, and minority ethnic) communities [145–147]. As pointed out by [148], hesitancy within this group is in part attributed to factors stemming from their exclusion in clinical vaccine trials [148]. Within this context, however, few studies have examined Asians as an independent subgroup within the broader framework of BAME research. [149], who focused on the Asian population as a distinct group in the US, found lower hesitancy compared to Black and Hispanic groups. Similarly, a national survey on COVID-19 vaccine intent among US racial and ethnic groups by [150] revealed that Asian Americans exhibited the lowest refusal rate (11%), in contrast to Black African Americans (32%) and American Indian/Alaska Native respondents (29%). Further research by [151], concentrating on ethnic minorities in a longitudinal study of UK households, found notably varying levels of vaccine hesitancy within Black and Pakistani/Bangladeshi ethnic groups. Exploring the context of the Black population, [152] indicated that higher rates of vaccine hesitancy among Black African Americans. According to [146], this is in part attributed to a lack of trust in medical institutions and concerns about racial injustice.

Turning now towards education, our findings show a general agreement between educational attainment and hesitancy across models. Specifically, they show how even a high school education can influence vaccine indecision, while some college education emerges as the most notable factor in most models. These findings support other related work in this area. For instance, [153] demonstrated heightened hesitancy within less-educated communities in high-income countries. Similarly, [112] observed that individuals with a high school level of education or lower were more inclined to exhibit vaccine hesitancy. Other work by [154], in a survey involving parents in Utah, US, noted that a significant portion of hesitant parents were from the middle-class and possessed either some college education or a college degree. Furthermore, several studies have corroborated that individuals with higher education levels or greater affluence tend to exhibit higher levels of vaccine hesitancy or even a refusal to be vaccinated altogether (e.g., [155–157]).

As it relates to individual household determinants, those with cohabitating couples also emerged as a significant factor to the hesitant or unsure, and strongly hesitant models respectively. Despite the limited amount of research in this area, most of the existing work show a decrease in hesitance rates with cohabilitation (e.g., [158–161]. One exception to this is the work of [162], which in addition to demonstrating increased vaccine hesitancy with

cohabilitation, also found being a member of the black population and having less than a college education to be important factors. Additionally, research conducted by [163], focusing on pregnant women in California, indicated that women with partners classified as essential workers exhibit a higher likelihood of hesitancy. The authors of this study propose that essential workers may have been previously exposed to the disease, leading respondents to believe they may have already developed immunity against COVID-19.

Another noteworthy factor gathered from our findings was occupation. The majority of variables within this category are associated with roles that support the day-to-day operations of various businesses, based on occupational classification data provided by the US Bureau of Labor Statistics [91]. Across most models, these trends manifest within three overarching occupational groups: Manufacturing, Professional, scientific, and management, and administrative and waste management services, and Wholesale trade. Work on occupation by [164] have reported more hesitant populations among older workers (aged 40 to 59) in sectors such as service and manufacturing, along with those who are unemployed. Similarly, [165], in a national survey encompassing Japanese adults aged 20 and older, observed reduced likelihood of vaccination among scientists and researchers. Temporal changes in hesitancy across different professions have also been explored in research. For instance, [166] examined shifts in hesitancy rates among various professions between January to May, and April to May 2021. The study highlighted substantial increases in categories such as computer/mathematical (7.3%), educators (9.0%), healthcare practitioners/technicians, and construction/extraction (45.2%) professionals. Furthermore, investigations have underscored variations in hesitancy levels within different healthcare groups [167]. It's important to note, however, that much of the existing work on vaccine hesitancy and occupation primarily focuses on roles within the medical sector, often classified as essential workers. Nevertheless, there remains limited research in this field [164]. On a related note, being employed was exclusively applicable to the probably not model. These results are corroborated by the findings of other work such as [48, 152]. However, there has been mixed findings in this respect (e.g., [168]) with further research needed to explore this issue.

Furthermore, income was shown to be an important factor, and in particular, per capita income, which displayed significance across all models. Previous investigations into per capita income have reported higher vaccine hesitancy among lower-income groups in comparison to their higher-income counterparts [169, 170]. Additionally, [171], in a survey focusing on parents with children aged 2 to 18, discovered that households with incomes under $100,000 exhibited lower vaccination likelihood than those with incomes of at least $150,000. In a related work, [172] identified families with both low education and income as reporting reduced willingness to vaccinate their children. Moreover, [173] established that the odds of vaccine hesitancy were twice as likely among individuals with middle income compared to those with lower income. Further, [174] analyzing households pre- and post-pandemic found that those who had an income of $100,000 or more prior to the pandemic and experienced income loss during the pandemic displayed heightened levels of hesitancy compared to those who didn't face income loss. As for the income range of $25,000 to $50,000 annually, these were only applicable to the hesitant and hesitant or unsure models, respectively.

Political leaning was a factor applicable to only three models: hesitant, unsure, and hesitant or unsure, affecting the the same groups. These inclinations encompassed affiliations with the Democratic, Green, Libertarian, and other political parties, which tended to be of lesser prevalence. Research into vaccine hesitancy has delved into these political associations. The majority of studies indicate that Democrats exhibit lower levels of vaccine hesitancy compared to their Republican counterparts. For instance, a study by [175] conducted on Americans revealed that 90% of surveyed Democrats had been vaccinated, while 68% of Independents and 58% of

Republicans displayed more hesitancy. Correspondingly, [135] found that Republican members were more inclined to oppose COVID-19 vaccination compared to Democrats. Additional research by [176] suggests that individuals with conservative views are less likely to trust scientific and medical experts, demonstrating a greater inclination to perceive vaccines as unsafe [177] and as a significant health threat [178]. These findings align with the conclusions drawn by [179], which identified that Conservatives and Republicans exhibited higher hesitancy levels compared to their Libertarian and Democrat counterparts.

The final set of variables concerns health, the level of preparedness, and the degree of concern regarding vaccine distribution at various locations. Most of these variables specifically apply to the uncertain model, while all models acknowledge the significance of the level of concern. Previous research has reported an association between vaccine uptake in hesitant communities and their level of apprehension regarding the accessibility and distribution of vaccines [111, 112]. To this end, regions facing greater hurdles in distributing vaccines, as gauged by the multidimensional CVAC index of concern, tend to exhibit lower vaccination rates compared to regions with fewer obstacles [180]. In terms of variables related to the number of COVID-19 cases and primary vaccinations, these findings are relevant to the hesitant and unsure models. Furthermore, the unsure model identifies the completion of a vaccination series as a notable factor. Conversely, social vulnerability is only pertinent to the uncertain model.

The above findings emphasize the complex interplay of diverse variables in shaping vaccine hesitancy. Factors such as education, ethnicity, age, and political orientation were crucial determinants across multiple models. These results underscore the importance of comprehensively understanding these factors to develop effective strategies for addressing vaccine hesitancy and promoting widespread vaccination. However, while this study focused on COVID-19 vaccines, the findings might not be generalizable to other vaccines for a variety of reasons. These concerns may stem from various factors such as their novelty, the short timeline for clinical trials, and the utilization of new mRNA technology in some of them.

With this being said, our work provides several valuable contributions. First, while the addition of the survey data resulted in greater model performance in comparison to the social media data, the speed and scale at which social media data can be collected and analysed makes it a supplementary source of data/insights on vaccine hesitancy. Second, and related, the results highlight limitations in using social-based data alone in understanding vaccine hesitancy. In this case, with the addition of the social media or survey data to the social-base data only models; this resulted in improved performances, as much as 17%. Third, there is an important link with respect to what is discussed online in cyber-social communities and the characteristics of the people in the real world. Finally, our approach lays the groundwork for other similar studies that seek to understand and compare the use of social media and survey data for other topics such as climate change and the economy.

There were also several limitations identified in this work that provide areas of future work. First, only 10 MSAs were examined in this study. Future work should therefore examine additional areas and at greater levels of spatial granularity. Second, working with text data is challenging, with the potential for different people to interpret the hesitance within text differently. In this study we did not assess the quality/agreement of the labelled data, which could be a source of bias in the results. Third, additional sources of social media [181] and survey data should be investigated, along with different machine learning algorithms [61, 72]). Fourth, as has been investigated in other research, other metrics derived from social media, such as sentiment (e.g., [61]) and stance (e.g., [51]), could be incorporated in a similar analysis. Fifth, this study exclusively used English tweets; a similar study encompassing multiple languages would therefore be interesting. Sixth, extending this study to encompass different countries would

yield valuable insights, considering the potential variations in hesitancy dynamics across diverse cultural and societal contexts.

Finally, it's important to note that individuals' decision to vaccinate or not is influenced by various factors, including individual capabilities, external opportunities, motivations, beliefs about necessity and concerns, and perceptions of health threats. Future work should therefore explore these results in the context of related frameworks, such as the Capability, Opportunity, Motivation, and Behavior Framework [182], the Necessity-Concerns Framework [183, 184], and the Health Belief Model [185], to gain a deeper understanding of the complex factors influencing vaccine hesitancy and inform targeted interventions to address this critical public health issue. Even with these areas for further exploration, this paper demonstrates the utility of utilizing both surveys and social media data in understanding vaccine hesitancy across different locations.

Author Contributions

Conceptualization: Kuleen Sasse, Ron Mahabir, Olga Gkountouna, Andrew Crooks, Arie Croitoru.

Data curation: Kuleen Sasse.

Formal analysis: Kuleen Sasse.

Investigation: Ron Mahabir, Olga Gkountouna.

Methodology: Ron Mahabir, Olga Gkountouna, Andrew Crooks, Arie Croitoru.

Supervision: Ron Mahabir, Olga Gkountouna.

Writing – original draft: Kuleen Sasse, Ron Mahabir, Olga Gkountouna.

Writing – review & editing: Ron Mahabir, Olga Gkountouna, Andrew Crooks, Arie Croitoru.

References

1. Dinleyici EC, Borrow R, Safadi MAP, van Damme P, Munoz FM. Vaccines and routine immunization strategies during the COVID-19 pandemic. Human vaccines & immunotherapeutics. 2021; 17(2):400–407. https://doi.org/10.1080/21645515.2020.1804776 PMID: 32845739
2. Gao Q, Bao L, Mao H, Wang L, Xu K, Yang M, et al. Development of an inactivated vaccine candidate for SARS-CoV-2. Science. 2020; 369(6499):77–81. https://doi.org/10.1126/science.abc1932 PMID: 32376603
3. Mohamed K, Rzymski P, Islam MS, Makuku R, Mushtaq A, Khan A, et al. COVID-19 vaccinations: The unknowns, challenges, and hopes. Journal of medical virology. 2022; 94(4):1336–1349. https://doi.org/10.1002/jmv.27487 PMID: 34845731
4. Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Gavrilov D, Giattino C, Hasell J, Macdonald B, Dattani S, Beltekian D, Ortiz-Ospina E, Roser M. Coronavirus (COVID-19) vaccinations Our World in Data. https://ourworldindata.org/covid-vaccinations.
5. Duca LM, Xu SF Likangand Price, McLean CA. Covid-19 stats: Covid-19 incidence, by age group—United States, March 1–November 14, 2020; 2020. Available from: https://www.cdc.gov/mmwr/volumes/69/wr/mm695152a8.htm.
6. DuRose R. The US has a new covid-19 strain on the rise. meet eris.; 2023. Available from: https://www.vox.com/health/2023/8/9/23825133/covid-eris-variant-eg5-cdc-pandemic-coronavirus-precautions-masking-booster-vaccine.
7. Irfan U. Covid-19 cases are exploding in Asia. here's what it means for the rest of the world.; 2022. Available from: https://www.vox.com/22977354/covid-19-outbreak-omicron-ba2-hong-kong-south-korea-china-asia-vaccine.
8. MacIntyre CR, Costantino V, Trent M. Modelling of COVID-19 vaccination strategies and herd immunity, in scenarios of limited and full vaccine supply in NSW, Australia. Vaccine. 2022; 40(17):2506–2513. https://doi.org/10.1016/j.vaccine.2021.04.042 PMID: 33958223

9. Morens DM, Folkers GK, Fauci AS. The concept of classical herd immunity may not apply to COVID-19. The Journal of Infectious Diseases. 2022; 226(2):195–198. https://doi.org/10.1093/infdis/jiac109 PMID: 35356987

10. Nohynek H, Wilder-Smith A. Does the world still need new Covid-19 vaccines? New England Journal of Medicine. 2022; 386(22):2140–2142. https://doi.org/10.1056/NEJMe2204695 PMID: 35507476

11. Holder J. Tracking coronavirus vaccinations around the world; 2023. Available from: https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html.

12. Johns Hopkins University & Medicine. Understanding vaccination progress; 2023. Available from: https://coronavirus.jhu.edu/vaccines/us-states#vaccination-rollout-us.

13. MacDonald NE, et al. Vaccine hesitancy: Definition, scope and determinants. Vaccine. 2015; 33 (34):4161–4164. https://doi.org/10.1016/j.vaccine.2015.04.036 PMID: 25896383

14. Troiano G, Nardi A. Vaccine hesitancy in the era of COVID-19. Public health. 2021; 194:245–251. https://doi.org/10.1016/j.puhe.2021.02.025 PMID: 33965796

15. Pierri F, Perry BL, DeVerna MR, Yang KC, Flammini A, Menczer F, et al. Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. Scientific reports. 2022; 12(1):5966. https://doi.org/10.1038/s41598-022-10070-w PMID: 35474313

16. Amin AB, Bednarczyk RA, Ray CE, Melchiori KJ, Graham J, Huntsinger JR, et al. Association of moral values with vaccine hesitancy. Nature Human Behaviour. 2017; 1(12):873–880. https://doi.org/10.1038/s41562-017-0256-5 PMID: 31024188

17. Dubé E, Gagnon D, Nickels E, Jeram S, Schuster M. Mapping vaccine hesitancy—Country-specific characteristics of a global phenomenon. Vaccine. 2014; 32(49):6649–6654. https://doi.org/10.1016/j.vaccine.2014.09.039 PMID: 25280436

18. Soares P, Rocha JV, Moniz M, Gama A, Laires PA, Pedro AR, et al. Factors associated with COVID-19 vaccine hesitancy. Vaccines. 2021; 9(3):300. https://doi.org/10.3390/vaccines9030300 PMID: 33810131

19. Hoy C, Wood T, Moscoe E. Addressing vaccine hesitancy in developing countries: Survey and experimental evidence. PLoS One. 2022; 17(11):e0277493. https://doi.org/10.1371/journal.pone.0277493 PMID: 36395260

20. Dixon S. Number of worldwide social network users 2027; 2023. Available from: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

21. Madziva R, Nachipo B, Musuka G, Chitungo I, Murewanhema G, Phiri B, et al. The role of social media during the COVID-19 pandemic: Salvaging its 'power'for positive social behaviour change in Africa. Health Promotion Perspectives. 2022; 12(1):22. https://doi.org/10.34172/hpp.2022.03 PMID: 35854855

22. Dixon SJ. Twitter: Annual growth rate worldwide 2024; 2022. Available from: https://www.statista.com/statistics/303723/twitters-annual-growth-rate-worldwide/.

23. Dixon S. Facebook: Annual growth rate worldwide 2024; 2022. Available from: https://www.statista.com/statistics/1202765/facebook-annual-growth-rate-worldwide/.

24. Mankala S, Durai A, Padiyar A, Gkountouna O, Mahabir R. Understanding public discourse surrounding the impact of bitcoin on the environment in social media. GeoJournal. 2023; p. 1–25. https://doi.org/10.1007/s10708-023-10856-z PMID: 38625109

25. Gunaratne K, Coomes EA, Haghbayan H. Temporal trends in anti-vaccine discourse on Twitter. Vaccine. 2019; 37(35):4867–4871. https://doi.org/10.1016/j.vaccine.2019.06.086 PMID: 31300292

26. Muric G, Wu Y, Ferrara E. COVID-19 vaccine hesitancy on social media: building a public Twitter data set of antivaccine content, vaccine misinformation, and conspiracies. JMIR public health and surveillance. 2021; 7(11):e30642. https://doi.org/10.2196/30642 PMID: 34653016

27. Sharma K, Zhang Y, Liu Y. COVID-19 vaccines: characterizing misinformation campaigns and vaccine hesitancy on twitter. arXiv preprint arXiv:210608423. 2021.

28. Ruiz J, Featherstone JD, Barnett GA. Identifying vaccine hesitant communities on twitter and their geolocations: a network approach. In: Proceedings of the 54th Hawaii international conference on system sciences; 2021. p. 3964.

29. Malova E. Understanding online conversations about covid-19 vaccine on twitter: vaccine hesitancy amid the public health crisis. Communication Research Reports. 2021; 38(5):346–356. https://doi.org/10.1080/08824096.2021.1983424

30. Piedrahita-Valdés H, Piedrahita-Castillo D, Bermejo-Higuera J, Guillem-Saiz P, Bermejo-Higuera JR, Guillem-Saiz J, et al. Vaccine hesitancy on social media: Sentiment analysis from June 2011 to April 2019. Vaccines. 2021; 9(1):28. https://doi.org/10.3390/vaccines9010028 PMID: 33430428

31. Garcia K, Berton L. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. Applied soft computing. 2021; 101:107057. https://doi.org/10.1016/j.asoc.2020.107057 PMID: 33519326

32. Al-Regaiey KA, Alshamry WS, Alqarni RA, Albarrak MK, Alghoraiby RM, Alkadi DY, et al. Influence of social media on parents' attitudes towards vaccine administration. Human Vaccines & Immunotherapeutics. 2022; 18(1):1872340. https://doi.org/10.1080/21645515.2021.1872340 PMID: 33596388

33. Biswas MR, Ali H, Ali R, Shah Z. Influences of social media usage on public attitudes and behavior toward COVID-19 vaccine in the Arab world. Human Vaccines & Immunotherapeutics. 2022; 18 (5):2074205. https://doi.org/10.1080/21645515.2022.2074205 PMID: 35671370

34. Ahmad Rizal AR, Nordin SM, Ahmad WFW, Ahmad Khiri MJ, Hussin SH. How does social media influence people to get vaccinated? The elaboration likelihood model of a person's attitude and intention to get COVID-19 vaccines. International Journal of Environmental Research and Public Health. 2022; 19 (4):2378. https://doi.org/10.3390/ijerph19042378 PMID: 35206563

35. Liew TM, Lee CS. Examining the utility of social media in COVID-19 vaccination: unsupervised learning of 672,133 twitter posts. JMIR public health and surveillance. 2021; 7(11):e29789. https://doi.org/10.2196/29789 PMID: 34583316

36. Yadav H, Sagar M. Exploring COVID-19 vaccine hesitancy and behavioral themes using social media big-data: a text mining approach. Kybernetes. 2023;. https://doi.org/10.1108/K-06-2022-0810

37. Piret J, Boivin G. Pandemics throughout history. Frontiers in microbiology. 2021; 11:631736. https://doi.org/10.3389/fmicb.2020.631736 PMID: 33584597

38. Nyawa S, Tchuente D, Fosso-Wamba S. COVID-19 vaccine hesitancy: a social media analysis using deep learning. Annals of Operations Research. 2022; p. 1–39. https://doi.org/10.1007/s10479-022-04792-3 PMID: 35729983

39. Larson HJ, Gakidou E, Murray CJ. The vaccine-hesitant moment. New England Journal of Medicine. 2022; 387(1):58–65. https://doi.org/10.1056/NEJMra2106441 PMID: 35767527

40. Machingaidze S, Wiysonge CS. Understanding COVID-19 vaccine hesitancy. Nature Medicine. 2021; 27(8):1338–1339. https://doi.org/10.1038/s41591-021-01459-7 PMID: 34272500

41. Pogue K, Jensen JL, Stancil CK, Ferguson DG, Hughes SJ, Mello EJ, et al. Influences on attitudes regarding potential COVID-19 vaccination in the United States. Vaccines. 2020; 8(4):582. https://doi.org/10.3390/vaccines8040582 PMID: 33022917

42. Haque A, Pant AB. Mitigating Covid-19 in the face of emerging virus variants, breakthrough infections and vaccine hesitancy. Journal of autoimmunity. 2022; p. 102792. https://doi.org/10.1016/j.jaut.2021.102792 PMID: 34995958

43. Boyon N, Silverstein K. Three in four adults globally say they would get a vaccine for COVID-19. Ipsos, News & Pools. 2020;.

44. Kreps S, Prasad S, Brownstein JS, Hswen Y, Garibaldi BT, Zhang B, et al. Factors associated with US adults' likelihood of accepting COVID-19 vaccination. JAMA network open. 2020; 3(10):e2025594–e2025594. https://doi.org/10.1001/jamanetworkopen.2020.25594 PMID: 33079199

45. Küçükali H, Ataç Ö, Palteki AS, Tokaç AZ, Hayran O. Vaccine hesitancy and anti-vaccination attitudes during the start of COVID-19 vaccination program: A content analysis on Twitter data. Vaccines. 2022; 10(2):161. https://doi.org/10.3390/vaccines10020161 PMID: 35214620

46. Shacham M, Greenblatt-Kimron L, Hamama-Raz Y, Martin LR, Peleg O, Ben-Ezra M, et al. Increased COVID-19 vaccination hesitancy and health awareness amid COVID-19 vaccinations programs in Israel. International journal of environmental research and public health. 2021; 18(7):3804. https://doi.org/10.3390/ijerph18073804 PMID: 33917327

47. Sallam M. COVID-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. Vaccines. 2021; 9(2):160. https://doi.org/10.3390/vaccines9020160 PMID: 33669441

48. Dror AA, Eisenbach N, Taiber S, Morozov NG, Mizrachi M, Zigron A, et al. Vaccine hesitancy: the next challenge in the fight against COVID-19. European journal of epidemiology. 2020; 35(8):775–779. https://doi.org/10.1007/s10654-020-00671-y PMID: 32785815

49. Downs JS, de Bruin WB, Fischhoff B. Parents' vaccination comprehension and decisions. In: Risk Analysis and Human Behavior. Routledge; 2013. p. 274–297.

50. Nyhan B, Reifler J, Richey S, Freed GL. Effective messages in vaccine promotion: a randomized trial. Pediatrics. 2014; 133(4):e835–e842. https://doi.org/10.1542/peds.2013-2365 PMID: 24590751

51. Nowak SA, Chen C, Parker AM, Gidengil CA, Matthews LJ. Comparing covariation among vaccine hesitancy and broader beliefs within Twitter and survey data. PloS one. 2020; 15(10):e0239826. https://doi.org/10.1371/journal.pone.0239826 PMID: 33031405

52. Karanja I. An enumeration and mapping of informal settlements in Kisumu, Kenya, implemented by their inhabitants. Environment and Urbanization. 2010; 22(1):217–239. https://doi.org/10.1177/0956247809362642

53. Mahfoud Z, Ghandour L, Ghandour B, Mokdad AH, Sibai AM. Cell phone and face-to-face interview responses in population-based surveys: how do they compare? Field methods. 2015; 27(1):39–54. https://doi.org/10.1177/1525822X14540084

54. Coffey SM, Elliott MR. Optimizing data collection interventions to balance cost and quality in a sequential multimode survey. Journal of Survey Statistics and Methodology. 2023; p. smad007. https://doi.org/10.1093/jssam/smad007

55. Krosnick JA. Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied cognitive psychology. 1991; 5(3):213–236. https://doi.org/10.1002/acp.2350050305

56. Hill SJ, Roberts ME. Acquiescence bias inflates estimates of conspiratorial beliefs and political misperceptions. Political Analysis. 2021; p. 1–16.

57. Romanchuk K, Linthwaite B, Cox J, Park H, Dussault C, Basta NE, et al. Determinants of SARS-CoV-2 vaccine willingness among people incarcerated in 3 Canadian federal prisons: a cross-sectional study. Canadian Medical Association Open Access Journal. 2022; 10(4):E922–E929. https://doi.org/10.9778/cmajo.20210248 PMID: 36280247

58. Dillman DA. Mail and Internet surveys: The tailored design method–2007 Update with new Internet, visual, and mixed-mode guide. John Wiley & Sons; 2011.

59. Al Baghal T, Sloan L, Jessop C, Williams ML, Burnap P. Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. Social Science Computer Review. 2020; 38(5):517–532. https://doi.org/10.1177/0894439319828011

60. Buntain C, McGrath E, Golbeck J, LaFree G. Comparing Social Media and Traditional Surveys around the Boston Marathon Bombing. In: # Microposts; 2016. p. 34–41.

61. Melotte S, Kejriwal M. Predicting zip code-level vaccine hesitancy in US Metropolitan Areas using machine learning models on public tweets. PLOS Digital Health. 2022; 1(4):e0000021. https://doi.org/10.1371/journal.pdig.0000021 PMID: 36812517

62. Swigger N. The online citizen: Is social media changing citizens' beliefs about democratic values? Political behavior. 2013; 35(3):589–603. https://doi.org/10.1007/s11109-012-9208-y

63. Smith PB, Fischer R. Acquiescence, extreme response bias and culture: A multilevel analysis. In: Multilevel analysis of individuals and cultures. Psychology Press; 2015. p. 285–314.

64. Joinson A. Social desirability, anonymity, and Internet-based questionnaires. Behavior Research Methods, Instruments, & Computers. 1999; 31(3):433–438. https://doi.org/10.3758/BF03200723 PMID: 10502866

65. Peddinti ST, Ross KW, Cappos J. User anonymity on twitter. IEEE Security & Privacy. 2017; 15 (3):84–87. https://doi.org/10.1109/MSP.2017.74

66. Pandya A, Lodha P. Social connectedness, excessive screen time during COVID-19 and mental health: a review of current evidence. Frontiers in Human Dynamics. 2021; p. 45.

67. Mwanza S, Suleman H. Measuring network structure metrics as a proxy for socio-political activity in social media. In: 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE; 2017. p. 878–883.

68. Baj-Rogowska A. Mapping of the Covid-19 vaccine uptake determinants from mining Twitter data. Ieee Access. 2021; 9:134929–134944. https://doi.org/10.1109/ACCESS.2021.3115554 PMID: 34786320

69. Nuzhath T, Tasnim S, Sanjwal RK, Trisha NF, Rahman M, Mahmud SF, et al. COVID-19 vaccination hesitancy, misinformation and conspiracy theories on social media: A content analysis of Twitter data; 2020.

70. Ma P, Zeng-Treitler Q, Nelson SJ. Use of two topic modeling methods to investigate covid vaccine hesitancy. In: Int. Conf. ICT Soc. Hum. Beings. vol. 384; 2021. p. 221–226.

71. Melton CA, Olusanya OA, Ammar N, Shaban-Nejad A. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. Journal of Infection and Public Health. 2021; 14(10):1505–1512. https://doi.org/10.1016/j.jiph.2021.08.010 PMID: 34426095

72. Qorib M, Oladunni T, Denis M, Ososanya E, Cotae P. COVID-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Systems with Applications. 2023; 212:118715. https://doi.org/10.1016/j.eswa.2022.118715 PMID: 36092862

73. Porreca A, Scozzari F, Di Nicola M. Using text mining and sentiment analysis to analyse YouTube Italian videos concerning vaccination. BMC Public Health. 2020; 20(1):1–9. https://doi.org/10.1186/s12889-020-8342-4 PMID: 32075631

74. Marcec R, Likic R. Using twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. Postgraduate medical journal. 2022; 98(1161):544–550. https://doi.org/10.1136/postgradmedj-2021-140685 PMID: 34373343

75. Chen N, Chen X, Pang J, Borga LG, D'Ambrosio C, Vögele C. Measuring COVID-19 Vaccine Hesitancy: Consistency of Social Media with Surveys. In: International Conference on Social Informatics. Springer; 2022. p. 196–210.

76. Peters MD. Addressing vaccine hesitancy and resistance for COVID-19 vaccines. International Journal of Nursing Studies. 2022; p. 104241. https://doi.org/10.1016/j.ijnurstu.2022.104241 PMID: 35489108

77. Savoia E, Piltch-Loeb R, Goldberg B, Miller-Idriss C, Hughes B, Montrond A, et al. Predictors of COVID-19 vaccine hesitancy: socio-demographics, co-morbidity, and past experience of racial discrimination. Vaccines. 2021; 9(7):767. https://doi.org/10.3390/vaccines9070767 PMID: 34358184

78. Ren J, Wagner AL, Zheng A, Sun X, Boulton ML, Huang Z, et al. The demographics of vaccine hesitancy in Shanghai, China. PLoS One. 2018; 13(12):e0209117. https://doi.org/10.1371/journal.pone.0209117 PMID: 30543712

79. Al-Wutayd O, Khalil R, Rajar AB. Sociodemographic and behavioral predictors of COVID-19 vaccine hesitancy in Pakistan. Journal of Multidisciplinary Healthcare. 2021; p. 2847–2856. https://doi.org/10.2147/JMDH.S325529 PMID: 34675532

80. Sabareesha SSS, Bhattacharjee S, Shetty RD. Pattern Analysis of COVID-19 Based On Geotagged Social Media Data with Sociodemographic Factors. In: 2022 27th International Conference on Automation and Computing (ICAC). IEEE; 2022. p. 1–6.

81. Lee J, Huang Y. COVID-19 Vaccine hesitancy: the role of socioeconomic factors and spatial effects. Vaccines. 2022; 10(3):352. https://doi.org/10.3390/vaccines10030352 PMID: 35334984

82. Tsou MH, Xu J, Lin CD, Daniels M, Embury J, Park J, et al. Analyzing Spatial-Temporal Impacts of Neighborhood Socioeconomic Status Variables on COVID-19 Outbreaks as Potential Social Determinants of Health. Annals of the American Association of Geographers. 2022; p. 1–22.

83. Brindley P, Cameron RW, Ersoy E, Jorgensen A, Maheswaran R. Is more always better? Exploring field survey and social media indicators of quality of urban greenspace, in relation to health. Urban Forestry & Urban Greening. 2019; 39:45–54. https://doi.org/10.1016/j.ufug.2019.01.015

84. Conrad FG, Gagnon-Bartsch JA, Ferg RA, Schober MF, Pasek J, Hou E. Social media as an alternative to surveys of opinions about the economy. Social Science Computer Review. 2021; 39(4):489–508. https://doi.org/10.1177/0894439319875692

85. Najafi H, Miller D. Comparing analysis of social media content with traditional survey methods of predicting opening night box-office revenues for motion pictures. Journal of Digital & Social Media Marketing. 2015; 3(3):262–278.

86. Google;. Available from: https://scholar.google.com/.

87. Clarivate;. Available from: https://www.webofscience.com.

88. Elsevier. Scopus;. Available from: https://www.webofscience.com.

89. Ventures S. Surgo Precision For Covid: U.S. COVID-19 vaccine coverage index;. Available from: https://vaccine.precisionforcovid.org/.

90. Bureau UC. American Community Survey; 2023. Available from: https://www.census.gov/programs-surveys/acs.

91. US Department of Labor. US Bureau of Labor Statistics; 2023. Available from: https://www.bls.gov/.

92. of Economic Analysis USB. Bureau of Economic Analysis;. Available from: https://www.bea.gov/.

93. Singer F. US 2020 General Official Election Results; 2021. Available from: https://par.nsf.gov/biblio/10312583.

94. Nytimes TNYT. Nytimes COVID-19-data;. Available from: https://github.com/nytimes/covid-19-data.

95. JHU. Understanding vaccination progress by country;. Available from: https://coronavirus.jhu.edu/vaccines/international.

96. Agency for Toxic Substances and Disease Registry (ATSDR). CDC/ATSDR Social Vulnerability Index; 2021. Available from: https://www.atsdr.cdc.gov/placeandhealth/svi/index.html.

97. HHS. Department of Health & Human Services; 2023. Available from: https://www.hhs.gov/.

98. CMU. Delphi;. Available from: https://vaccine-hesitancy.healthdata.org.

99. Yuan X, Schuchard RJ, Crooks AT. Examining emergent communities and social bots within the polarized online vaccination debate in Twitter. Social media + society. 2019; 5(3):2056305119865465.

100. ndougla6. Hesitant-tweet-discussion: Files related to hesitant tweet paper, GitHub;. Available from: https://github.com/ndougla6/hesitant-tweet-discussion.

101. Chen Q, Crooks A. Analyzing the vaccination debate in social media data Pre-and Post-COVID-19 pandemic. International Journal of Applied Earth Observation and Geoinformation. 2022; 110:102783. https://doi.org/10.1016/j.jag.2022.102783 PMID: 35528967

102. Epidemiology D. Digitalepidemiologylab/multi-lang-vaccine-sentiment: Different approaches to multi-language vaccine sentiment models;. Available from: https://github.com/digitalepidemiologylab/multi-lang-vaccine-sentiment.

103. Twitter. Twitter API;. Available from: https://developer.twitter.com/en/docs/twitter-api.

104. Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL, et al. The measles vaccination narrative in Twitter: a quantitative analysis. JMIR public health and surveillance. 2016; 2 (1):e5059. https://doi.org/10.2196/publichealth.5059 PMID: 27227144

105. Chen Q, Croitoru A, Crooks A. A comparison between online social media discussions and vaccination rates: A tale of four vaccines Digital Health. 2023; 9:102783. https://doi.org/10.1177/20552076231155682

106. Cascini F, Pantovic A, Al-Ajlouni YA, Failla G, Puleo V, Melnyk A, Lontano A, Ricciardi W. Social media and attitudes towards a COVID-19 vaccination: A systematic review of the literature EClinical-Medicine. 2022; 48:101454. https://doi.org/10.1016/j.eclinm.2022.101454 PMID: 35611343

107. Mackun P, Wilson S, Fischetti T, et al. Population distribution and change: 2000 to 2010: US Department of Commerce. Economics and Statistics Administration, US Census Bureau. 2011;.

108. Koiranen I, Keipi T, Koivula A, Räsänen P. Changing patterns of social media use? A population-level study of Finland. Universal Access in the Information Society. 2020; 19:603–617. https://doi.org/10.1007/s10209-019-00654-1

109. Kulshrestha J, Kooti F, Nikravesh A, Gummadi K. Geographic dissection of the Twitter network. Proceedings of the International AAAI Conference on Web and Social Media. 2021; 6(1):202–209. https://doi.org/10.1609/icwsm.v6i1.14280

110. Blank G. The digital divide among Twitter users and its implications for social research. Social Science Computer Review. 2017; 35(6):679–697. https://doi.org/10.1177/0894439316671698

111. Tolley AJ, Scott VC, Mitsdarffer ML, Scaccia JP. The Moderating Effect of Vaccine Hesitancy on the Relationship between the COVID-19 Vaccine Coverage Index and Vaccine Coverage. Vaccines. 2023; 11(7):1231. https://doi.org/10.3390/vaccines11071231 PMID: 37515046

112. Khairat S, Zou B, Adler-Milstein J. Factors and reasons associated with low COVID-19 vaccine uptake among highly hesitant communities in the US. American journal of infection control. 2022; 50(3):262–267. https://doi.org/10.1016/j.ajic.2021.12.013 PMID: 34995722

113. Andersen CM, Bro R Variable selection in regression—a tutorial Journal of Chemometrics. 2010; 24 (11-12):728–737.

114. Keany E. Boruta Shap project; 2021. Available from: https://pypi.org/project/BorutaShap/.

115. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010; 36(11):1–13. https://doi.org/10.18637/jss.v036.i11

116. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 4765–4774. Available from: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

117. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence. 2020; 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9 PMID: 32607472

118. Keany E. Is this the best feature selection algorithm "borutashap"?; 2020. Available from: https://medium.com/analytics-vidhya/is-this-the-best-feature-selection-algorithm-borutashap-8bc238aa1677.

119. Ghosh I, Chaudhuri TD. Integrating Navier-Stokes equation and neoteric iForest-BorutaShap-Facebook's prophet framework for stock market prediction: An application in Indian context. Expert Systems with Applications. 2022; 210:118391. https://doi.org/10.1016/j.eswa.2022.118391

120. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017; 30.

121. Silva I, Ferreira C, Costa L, Sóter M, Carvalho L, de C Albuquerque J, et al. Polycystic ovary syndrome: clinical and laboratory variables related to new phenotypes using machine-learning models. Journal of Endocrinological Investigation. 2022; p. 1–9.

122. keany E. Boruta-Shap;. Available from: https://github.com/Ekeany/Boruta-Shap.

123. Chatterjee S, Krystyanczuk M. Python social media analytics. Packt Publishing Ltd; 2017.

124. Adebayo GO, Yampolskiy RV. Estimating intelligence quotient using stylometry and machine learning techniques: A review. Big Data Mining and Analytics. 2022; 5(3):163–191. https://doi.org/10.26599/BDMA.2022.9020002

125. Sivaram P, Senthilkumar S, Gupta L, Lokesh NS. Perspectives on Social Welfare Applications Optimization and Enhanced Computer Applications. IGI Global; 2023.

126. Nguyen DQ, Vu T, Nguyen AT. BERTweet: A pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; 2020. p. 9–14.

127. Nguyen DQ, Vu T, Nguyen AT. BERTweet: A pre-trained language model for English Tweets. arXiv preprint arXiv:200510200. 2020;.

128. Liaw R, Liang E, Nishihara R, Moritz P, Gonzalez JE, Stoica I. Tune: A Research Platform for Distributed Model Selection and Training. arXiv preprint arXiv:180705118. 2018;.

129. Obeidat R, Gharaibeh M, Abdullah M, Alharahsheh Y. Multi-label multi-class COVID-19 Arabic Twitter dataset with fine-grained misinformation and situational information annotations. PeerJ Computer Science. 2022; 8:e1151. https://doi.org/10.7717/peerj-cs.1151 PMID: 36532803

130. Ozyurt B, Akcayol MA. A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA. Expert Systems with Applications. 2021; 168:114231. https://doi.org/10.1016/j.eswa.2020.114231

131. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

132. Chen T, Guestrin C XGBoost: A Scalable Tree Boosting System KDD'16 Conference. 2016:785–794.

133. McKnight PE, Najab J. Mann-Whitney U Test. The Corsini encyclopedia of psychology. 2010; p. 1–1. https://doi.org/10.1002/9780470479216.corpsy0524

134. Sigalo N, Awasthi N, Abrar SM, Frias-Martinez V. Using COVID-19 vaccine attitudes on Twitter to improve vaccine uptake forecast models in the United States: infodemiology study of Tweets JMIR infodemiology. 2023; 3(1):e43703. https://doi.org/10.2196/43703 PMID: 37390402

135. Fridman A, Gershon R, Gneezy A. COVID-19 and vaccine hesitancy: A longitudinal study. PloS one. 2021; 16(4):e0250123. https://doi.org/10.1371/journal.pone.0250123 PMID: 33861765

136. Watson OJ, Barnsley G, Toor J, Hogan AB, Winskill P, Ghani AC. Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. The Lancet Infectious Diseases. 2022; 22 (9):1293–1302. https://doi.org/10.1016/S1473-3099(22)00320-6 PMID: 35753318

137. Steele MK, Couture A, Reed C, Iuliano D, Whitaker M, Fast H, et al. Estimated number of COVID-19 infections, hospitalizations, and deaths prevented among vaccinated persons in the US, December 2020 to September 2021. JAMA Network Open. 2022; 5(7):e2220385–e2220385. https://doi.org/10.1001/jamanetworkopen.2022.20385 PMID: 35793085

138. University B. Vaccinations; 2022. Available from: https://globalepidemics.org/vaccinations/.

139. Sedgwick P. Impact of vaccine hesitancy on secondary COVID-19 outbreaks in the US: an age-structured SIR model. BMJ. 344.

140. Mislove A, Lehmann S, Ahn Y, Onnela J, Rosenquist J Understanding the demographics of Twitter users Proceedings of the international AAAI conference on web and social media. 2011; 5(1):554–557. https://doi.org/10.1609/icwsm.v5i1.14168

141. Xiao Y, Huang Q, Wu K Understanding social media data for disaster management Natural hazards. 2015; 79:1663–1679. https://doi.org/10.1007/s11069-015-1918-0

142. Hong L, Convertino G, Chi E Language matters in twitter: A large scale study Proceedings of the international AAAI conference on web and social media. 2011; 5(1):518–521. https://doi.org/10.1609/icwsm.v5i1.14184

143. Murphy J, Vallières F, Bentall RP, Shevlin M, McBride O, Hartman TK, et al. Psychological characteristics associated with COVID-19 vaccine hesitancy and resistance in Ireland and the United Kingdom. Nature communications. 2021; 12(1):29. https://doi.org/10.1038/s41467-020-20226-9 PMID: 33397962

144. Fazel M, Puntis S, White SR, Townsend A, Mansfield KL, Viner R, et al. Willingness of children and adolescents to have a COVID-19 vaccination: results of a large whole schools survey in England. EClinicalMedicine. 2021; 40:101144. https://doi.org/10.1016/j.eclinm.2021.101144 PMID: 34608453

145. Willis DE, Andersen JA, Bryant-Moore K, Selig JP, Long CR, Felix HC, et al. COVID-19 vaccine hesitancy: Race/ethnicity, trust, and fear. Clinical and translational science. 2021; 14(6):2200–2207. https://doi.org/10.1111/cts.13077 PMID: 34213073

146. Padamsee TJ, Bond RM, Dixon GN, Hovick SR, Na K, Nisbet EC, et al. Changes in COVID-19 vaccine hesitancy among Black and White individuals in the US. JAMA Network Open. 2022; 5(1):e2144470–e2144470. https://doi.org/10.1001/jamanetworkopen.2021.44470 PMID: 35061038

147. Curtis MG, Whalen CC, Pjesivac I, Kogan SM. Contextual Pathways Linking Cumulative Experiences of Racial Discrimination to Black American Men's COVID Vaccine Hesitancy. Journal of Racial and

Ethnic Health Disparities. 2022; p. 1–13. https://doi.org/10.1007/s40615-022-01471-8 PMID: 36445684

148. Pepperrell T, Rodgers F, Tandon P, Sarsfield K, Pugh-Jones M, Rashid T, et al. Making a COVID-19 vaccine that works for everyone: ensuring equity and inclusivity in clinical trials. Global Health Action. 2021; 14(1):1892309. https://doi.org/10.1080/16549716.2021.1892309 PMID: 33627051

149. Nguyen LH, Joshi AD, Drew DA, Merino J, Ma W, Lo CH, et al. Self-reported COVID-19 vaccine hesitancy and uptake among participants from different racial and ethnic groups in the United States and United Kingdom. Nature communications. 2022; 13(1):636. https://doi.org/10.1038/s41467-022-28200-3 PMID: 35105869

150. Nápoles AM, Stewart AL, Strassle PD, Quintero S, Bonilla J, Alhomsi A, et al. Racial/ethnic disparities in intent to obtain a COVID-19 vaccine: A nationally representative United States survey. Preventive Medicine Reports. 2021; 24:101653. https://doi.org/10.1016/j.pmedr.2021.101653

151. Robertson E, Reeve KS, Niedzwiedz CL, Moore J, Blake M, Green M, et al. Predictors of COVID-19 vaccine hesitancy in the UK household longitudinal study. Brain, behavior, and immunity. 2021; 94:41–50. https://doi.org/10.1016/j.bbi.2021.03.008 PMID: 33713824

152. Gerretsen P, Kim J, Caravaggio F, Quilty L, Sanches M, Wells S, et al. Individual determinants of COVID-19 vaccine hesitancy. PloS one. 2021; 16(11):e0258462. https://doi.org/10.1371/journal.pone.0258462 PMID: 34788308

153. Bergen N, Kirkby K, Fuertes CV, Schlotheuber A, Menning L, Mac Feely S, et al. Global state of education-related inequality in COVID-19 vaccine coverage, structural barriers, vaccine hesitancy, and vaccine refusal: findings from the Global COVID-19 Trends and Impact Survey. The Lancet Global Health. 2023; 11(2):e207–e217. https://doi.org/10.1016/S2214-109X(22)00520-4 PMID: 36565702

154. Luthy KE, Beckstrand RL, Callister LC. Parental hesitation in immunizing children in Utah. Public Health Nursing. 2010; 27(1):25–31. https://doi.org/10.1111/j.1525-1446.2009.00823.x PMID: 20055965

155. Hegde S, Wagner A, Clarke P, Potter R, Swanson R, Boulton M. Neighbourhood influence on the fourth dose of diphtheria-tetanus-pertussis vaccination. Public Health. 2019; 167:41–49. https://doi.org/10.1016/j.puhe.2018.11.009 PMID: 30639802

156. Smith PJ, Humiston SG, Marcuse EK, Zhao Z, Dorell CG, Howes C, et al. Parental delay or refusal of vaccine doses, childhood vaccination coverage at 24 months of age, and the Health Belief Model. Public health reports. 2011; 126(2_suppl):135–146. https://doi.org/10.1177/00333549111260S215 PMID: 21812176

157. McNutt LA, Desemone C, DeNicola E, El Chebib H, Nadeau JA, Bednarczyk RA, et al. Affluence as a predictor of vaccine refusal and underimmunization in California private kindergartens. Vaccine. 2016; 34(14):1733–1738. https://doi.org/10.1016/j.vaccine.2015.11.063 PMID: 26679403

158. Thorneloe R, Wilcockson H, Lamb M, Jordan CH, Arden M. Willingness to receive a COVID-19 vaccine among adults at high-risk of COVID-19: a UK-wide survey; 2020.

159. Huang H, Zhu XM, Liang PW, Fang ZM, Luo W, Ma YM, et al. COVID-19 vaccine uptake, acceptance, and hesitancy among persons with mental disorders during the second stage of China's nationwide vaccine rollout. Frontiers in Medicine. 2021; 8:761601. https://doi.org/10.3389/fmed.2021.761601 PMID: 34901076

160. Jain A, Van Hoek AJ, Boccia D, Thomas SL. Lower vaccine uptake amongst older individuals living alone: A systematic review and meta-analysis of social determinants of vaccine uptake. Vaccine. 2017; 35(18):2315–2328. https://doi.org/10.1016/j.vaccine.2017.03.013 PMID: 28343775

161. Zakeri M, Li J, Sadeghi SD, Essien EJ, Sansgiry SS. Strategies to decrease COVID-19 vaccine hesitancy for children. Journal of Pharmaceutical Health Services Research. 2021; 12(4):539–544. https://doi.org/10.1093/jphsr/rmab060

162. Klinkhammer KE, Romm KF, Kerrigan D, McDonnell KA, Vyas A, Wang Y, et al. Sociopolitical, mental health, and sociodemographic correlates of COVID-19 vaccine hesitancy among young adults in 6 US metropolitan areas. Preventive medicine reports. 2022; 27:101812. https://doi.org/10.1016/j.pmedr.2022.101812 PMID: 35541217

163. Simmons LA, Whipps MD, Phipps JE, Satish NS, Swamy GK. Understanding COVID-19 vaccine uptake during pregnancy:'Hesitance', knowledge, and evidence-based decision-making. Vaccine. 2022; 40(19):2755–2760. https://doi.org/10.1016/j.vaccine.2022.03.044 PMID: 35361501

164. Takahashi S, Takahashi N, Sasaki S, Nohara M, Kawachi I. Occupational disparities in COVID-19 vaccine hesitancy in Japan. SSM-population health. 2022; 19:101226. https://doi.org/10.1016/j.ssmph.2022.101226 PMID: 36119724

165. Nomura S, Eguchi A, Yoneoka D, Kawashima T, Tanoue Y, Murakami M, et al. Reasons for being unsure or unwilling regarding intention to take COVID-19 vaccine among Japanese people: A large cross-sectional national survey. The Lancet Regional Health-Western Pacific. 2021; 14:100223. https://doi.org/10.1016/j.lanwpc.2021.100223 PMID: 34368797

166. King WC, Rubinstein M, Reinhart A, Mejia R. COVID-19 vaccine hesitancy January-May 2021 among 18–64 year old US adults by employment and occupation. Preventive medicine reports. 2021; 24:101569. https://doi.org/10.1016/j.pmedr.2021.101569 PMID: 34603943

167. Gatto NM, Lee JE, Massai D, Zamarripa S, Sasaninia B, Khurana D, et al. Correlates of COVID-19 vaccine acceptance, hesitancy and refusal among employees of a safety net California county health system with an early and aggressive vaccination program: results from a cross-sectional survey. Vaccines. 2021; 9(10):1152. https://doi.org/10.3390/vaccines9101152 PMID: 34696260

168. Marzo RR, Sami W, Alam MZ, Acharya S, Jermsittiparsert K, Songwathana K, et al. Hesitancy in COVID-19 vaccine uptake and its associated factors among the general adult population: a cross-sectional study in six Southeast Asian countries. Tropical Medicine and Health. 2022; 50:1–10. https://doi.org/10.1186/s41182-021-00393-1 PMID: 34983692

169. Guo Y, Kaniuka AR, Gao J, Sims OT. An epidemiologic analysis of associations between county-level per capita income, unemployment rate, and COVID-19 vaccination rates in the United States. International Journal of Environmental Research and Public Health. 2022; 19(3):1755. https://doi.org/10.3390/ijerph19031755 PMID: 35162778

170. Hughes MM, Wang A, Grossman MK, Pun E, Whiteman A, Deng L, et al. County-level COVID-19 vaccination coverage and social vulnerability—United States, December 14, 2020–March 1, 2021. Morbidity and Mortality Weekly Report. 2021; 70(12):431. https://doi.org/10.15585/mmwr.mm7012e1 PMID: 33764963

171. McKinnon B, Quach C, Dubé È, Nguyen CT, Zinszer K. Social inequalities in COVID-19 vaccine acceptance and uptake for children and adolescents in Montreal, Canada. Vaccine. 2021; 39 (49):7140–7145. https://doi.org/10.1016/j.vaccine.2021.10.077 PMID: 34763947

172. Hetherington E, Edwards SA, MacDonald SE, Racine N, Madigan S, McDonald S, et al. SARS-CoV-2 vaccination intentions among mothers of children aged 9 to 12 years: a survey of the All Our Families cohort. Canadian Medical Association Open Access Journal. 2021; 9(2):E548–E555.

173. Hassen HD, Welde M, Menebo MM. Understanding determinants of COVID-19 vaccine hesitancy; an emphasis on the role of religious affiliation and individual's reliance on traditional remedy. BMC Public Health. 2022; 22(1):1–11. https://doi.org/10.1186/s12889-022-13485-2

174. Jantzen R, Maltais M, Broët P. Socio-demographic factors associated with COVID-19 vaccine hesitancy among middle-aged adults during the Quebec's vaccination campaign. Frontiers in Public Health. 2022; p. 521. https://doi.org/10.3389/fpubh.2022.756037 PMID: 35372193

175. Hamel L, Lopes L, Sparks G, Stokes M, Brodie M. KFF COVID-19 vaccine monitor: April 2021. Kaiser Family Foundation. 2021; 6.

176. Motta M. The dynamics and political implications of anti-intellectualism in the United States. American Politics Research. 2018; 46(3):465–498. https://doi.org/10.1177/1532673X17719507

177. Baumgaertner B, Carlisle JE, Justwan F. The influence of political ideology and trust on willingness to vaccinate. PloS one. 2018; 13(1):e0191728. https://doi.org/10.1371/journal.pone.0191728 PMID: 29370265

178. Tyson A. Republicans remain far less likely than Democrats to view COVID-19 as a major threat to public health; 2020.

179. Tyson A, Funk C, Kennedy B, Johnson C. Majority in US says public health benefits of COVID-19 restrictions worth the costs, even as large shares also see downsides; 2021.

180. Mishra A, Sutermaster S, Smittenaar P, Stewart N, Sgaier SK. COVID-19 Vaccine Coverage Index: Identifying barriers to COVID-19 vaccine uptake across US counties. MedRxiv. 2021; p. 2021–06.

181. Maugeri A, Barchitta M, Agodi A Using google trends to predict COVID-19 vaccinations and monitor search behaviours about vaccines: A retrospective analysis of italian data Vaccines. 2022; 10(1):119. https://doi.org/10.3390/vaccines10010119 PMID: 35062780

182. Michie S, Van Stralen MM, West R The behaviour change wheel: a new method for characterising and designing behaviour change interventions Implementation science. 2021; 6(1):1–12.

183. Phillips LA, Diefenbach MA, Kronish IM, Negron RM, Horowitz CR The necessity-concerns framework: a multidimensional theory benefits from multidimensional analysis Implementation science. 2014; 48 (1):7–16. https://doi.org/10.1007/s12160-013-9579-2 PMID: 24500078

184. Jones CL, Jensen JD, Scherr CL, Brown NR, Christy K, Weaver J The health belief model as an explanatory framework in communication research: exploring parallel, serial, and moderated mediation Health communication. 2015; 30(6):566–576. https://doi.org/10.1080/10410236.2013.873363 PMID: 25010519

185. Horne R, Chapman SCE, Parham R, Freemantle N, Forbes A, Cooper V Understanding patients' adherence-related beliefs about medicines prescribed for long-term conditions: a meta-analytic review of the Necessity-Concerns Framework PloS one. 2013; 8(12):e80633. https://doi.org/10.1371/journal.pone.0080633 PMID: 24312488