Impacts of Misinformation on COVID-19 Vaccine Administration in the U.S.

CTSP Research Project

Author: Montoya, Zachary B.S. Ch.E.
Advisor: Dadashi, Andisheh M.S. Stat

Albuquerque, New Mexico

In association with

National Security Studies Program (NSSP)
Critical Technology Studies Program (CTSP)
The University of New Mexico
Computer Science Department
Albuquerque, New Mexico

The University of New Mexico – Valencia Campus
DATED: Summer 2021

Farris Engineering Center
901 Redondo South Dr. Albuquerque, NM 87106.

Abstract

The scope of this research concerns vaccines and online misinformation. In the recent decades, vaccines have been a high-profile topic of interest in the public's perception. Historically, the implementation of vaccines has been negatively impacted due to a lack of information or misinformation. The exposure and controversy surrounding them have appreciably increased in the last year due to the COVID-19 pandemic. Vaccine biotechnology has the potential to greatly improve the hosts immunological response to virus and public health through the concept of herd immunity. The objective of this document is to investigate the impact and dangers of misinformation on COVID-19 vaccination hesitancy. The relationship between incidents of vaccine related misinformation on Twitter and COVID-19 vaccine administration rates were evaluated for dependence and found to be unrelated within the United States during January 2021. This preliminary study has potential utility for further research on topic.

Introduction

In 2020 alone, the SARS-COV-2 virus that generates a respiratory illness in humans, rose to 10,719.92 confirmed infections per million people[1]. Bear in mind that the number of confirmed cases is less than the number of actual cases primarily due to testing being a limiting factor[2]. This has created an impact in nearly 220 territories and countries across the globe[3]. Given this, the precedent is set on how far a pandemic may spread with regards to transmissibility and sets forth an idea of the number of people it can potentially impact.

Historically humanity has experienced other pandemics on a global scale with massive impacts on life, such as the 1918 Spanish Flu Pandemic and HIV Global Pandemic[4], [5]. Although an estimated 50 million humans infected with the 1918 Spanish Flu during the pandemic died, this figure may be an underestimate and not accurately capture the real death toll from this historical event[6]. It is, however, with great certainty that diseases such as SARS, Influenza, and HIV will continue to evolve, and others will appear. They may develop through animal origins as approximately 75% of emerging diseases are zoonotic[7]. This escalates the potential for pandemics to continue in the future, as humans increasingly impinge on wildlife habitats. This risk of pandemic is also increased by critical parameters such as urbanization, high-speed global travel, and rising populations[8].

Scientists around the globe are working tirelessly in an effort to stay ahead of the epidemiology trends by investigating highly-probably zoonotic causes of infection[9] and conducting research and development on forecasted vaccine solutions[10]. Focusing on the former of the two aforementioned efforts, vaccines and inoculation technologies are a strong candidate for continued solutions for preventing the spread of diseases. Vaccines have experienced friction on a number of fronts. One of the most notable repercussions came from a case series authored by Andrew Wakefield's 1998, now retracted, that implied a relationship between the measles, mumps, and rubella (MMR) vaccine and juvenile developmental disorders and behavioral regression[11]. Wakefield's work had initiated response studies testing the logic and pathophysiology behind MMR causing these disorders. Investigation of work's findings then found Wakefield *et al.* guilty of intentional scientific fraud which led to retraction of the document in its entirety. [12]. The deception from this work became a sensation at the time and has since propagated ripple effects on vaccine acceptance. Despite objective evidence from reputable researchers and organizations, some people maintain concerns that autism spectrum disorder (ASD) is related to or caused by vaccines. ASD is a pervasive developmental disorder that causes significant regressive behavioral disorders[13]. There is a large pool of existing research that shows vaccines are not the cause of ASD. A 2013 Center for Disease Control (CDC) study expanded this pool by concluding antigens in inoculated children with and without ASD were the same[14]. An antigen is any compound that causes the human body's immune system to create antibodies in response, therefore it was a critical experiment output in this study providing equitable proof[15].

The sentiment initialized from misinformation *Wakefield et al.* provided, suggesting that vaccines can be inherently dangerous, is still standing but on a new stage in 2021. Misinformation and distrust enveloping the COVID-19 vaccines are promoting immunization tentativeness. Which then has the ability to poses a large threat to public health with respect to herd immunity[16]. Census survey data indicates that a disturbing percent of the population of American adults will “probably not” or “definitely not” receive the COVID-19 vaccination when it is available for them[17]. Contemporary researchers are seeking to illustrate a clearer picture of the impact that online misinformation has had on COVID-19 vaccinations. There are concerning results on the significance of the impact. *F. Pierri et. al.* showing a significant ($P<0.001$) negative association between the two parameters, suggesting that as misinformation decreases vaccinations increase[18]. Similarly, *V. Carrieri et. al.* shows that misinformation and fake news in certain geographical regions resulted in inoculation rates below the herd immunity convention threshold[19].

In this paper we ask, if misinformation in social media will impact COVID-19 immunization rates in the United States? In particular, how does the number of tweets posted in Twitter providing misinformation impact immunization rates?

Methods

The incidence rate of tweets containing misinformation pertinent to vaccines in the United States is the critical independent variable in this analysis. The dependent variables include inoculation rates (vaccine administration rates) within the U.S.

Tweets containing Covid-19 and vaccine related information were collected using the CoVaxxy Public Database collaboratively hosted by Indiana University and Texas Advanced Computing Center. Relevant content from Twitter is identified using a keyword snowball sampling method[20], that is capable of matching with substrings such as main content, URLs, and hashtags. Tweet data is sampled and filtered daily into this database beginning on January 5th, 2021, to present day. The volume of tweets meeting the relevance criterion in the filtering decreased vicinal January 6th, 2021, possibly due to the storming of the U.S. Capitol.

Twitter data policy dictates that only tweet IDs can be shared publicly, therefore tweets IDs were only collected using the CoVaxxy database[21]. The data was then *hydrated* to JSON files using the Twarc Python package to interface with the standard Twitter API v2. Due largely to the Twitter API v2 developer rate limits of 300 Tweets per 15-minute interval, the sample size in-scope of this study was selected to be January 5th, 2021, to January 31st, 2021[22]. This window was selected because on January 4th White House announced the controversial Operation Warp Speed (OWS), a project to accelerate the development, production, and inoculation of the COVID-19 vaccines. On January 6th, 2021, the CDC detailed that 21.4MM doses had been distributed and 5.9MM doses of the COVID-19 vaccine from the various manufacturers involved in OWS, had been administered to patients[23].

This selected sample size from January 5th to January 31st contains 11MM tweets. The JSON files were converted to CSV files using a Python script for simplified data handling. The CSV files were aggregated after hydration to determine their credibility. Tweets containing URLs with their location visible were filtered through from the initial data set, then domains were then evaluated for credibility against an index of 813 domains of known credibility, compiled by an independent third-party[24], equally used in research on misinformation[25]–[28]. This index ranks only domains on their credibility (on a scale from low to high), as ranking individual articles would not be feasible due to the sheer daily volume. (See S.I. for more information).

Data provided with in the Tweet JSON files was used to attribute the geolocation of the 1.8MM Tweets, the remaining tweets were unable to be accurately geolocated due to user privacy settings on Twitter. Daily COVID-19 vaccination distribution and administration data was collected from the CDC database and added on a state-by-state basis[29]. It should be noted throughout the duration of the studied time interval vaccine distribution values were greater than the values of vaccines administered. This may suggest that there was a greater supply than actual administered or that the inventory was in transit and could not be utilized.

Regression modeling was used to determine the relationship between the incidence rate of Tweets shared from low credibility sources and the COVID-19 vaccine administration rates. The framework of this model is as shown in the following multiple regression equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad \text{Eqn 1}$$

Where,

Y_i: is the dependent variable,

β_0: is the population Y intercept,

β_1 and β_2 are population slope coefficients,

X_{i1} and X_{i2} are independent variables, and

ε_i: is the random error term.

In the multiple regression model found in Table 2, X_{i1} and X_{i2}, are representative of the Date and Tweet Misinformation.

A variable inflation factor (VIF) factor was used to determine potential effects of multicollinearity between independent variables within the model. VIF was calculated using R statistical programming language. VIF is conventionally calculated using the following equation:

$$VIF_i = \frac{1}{1 - R_i^2} \qquad \text{Eqn 2}$$

$$for\ i = 1, 2, \ldots, k$$

Where, R_i^2 is the coefficient of multiple determination of X_i on the remaining explanatory variables[30].

Based on this equation, a VIF of 1 indicates that the independent variable regression coefficients are free of collinearity. Conversely, the larger the VIF value the greater the influence of collinearity between independent variable, meaning that variance in regression coefficients increase. Large variance in regression coefficients create unreliable probabilities and confidence intervals. According to conventional statistics and academic research, if VIF is greater than 5 or lower than 0.2, multicollinearity is considered to be present[31], [32].

If multi-linearity is not present, determined by a VIF less than 5 and ideally closer to 1, then confidence in the regression coefficients for the explanatory variables (independent variables) increases. The VIF for a multiple regression model can determine if multicollinearity is occurring. In a model with a large number of independent variables, it cannot determine the independent variables causing multi-collinearity.

In order to address multicollinearity within a model, the independent variables that are highly correlated can be removed from the set or a principal component analysis can be conducted to reduce data dimension through decomposition.

Results

The following figures and tables found on the subsequent pages capture two linear regression models and a multiple regression with a variable inflation factor analysis for the multiple regression model.

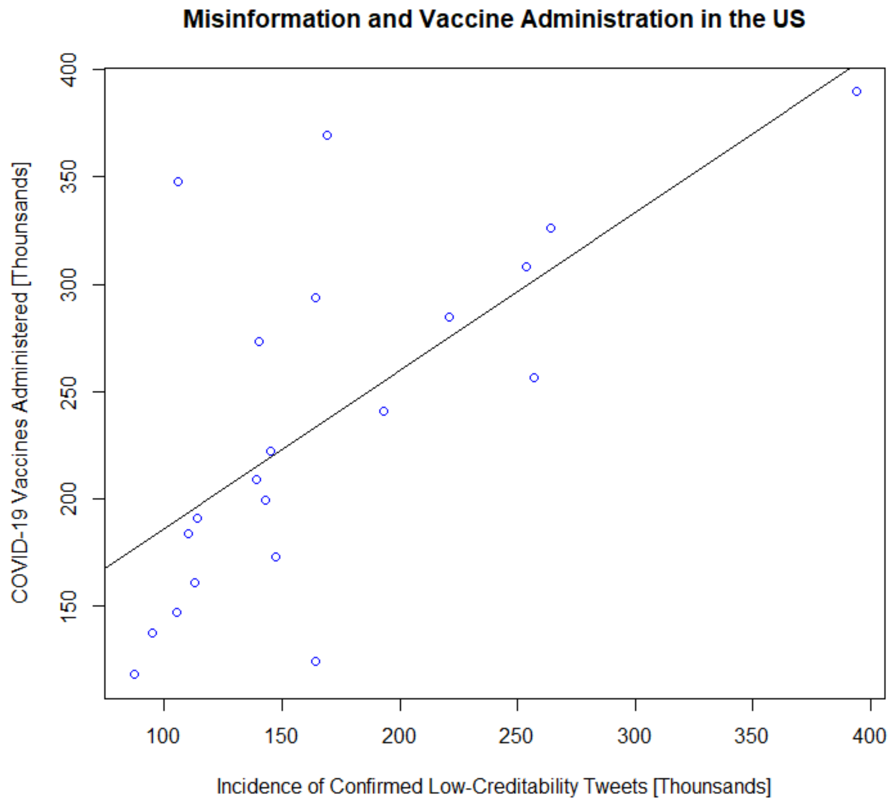

Figure 1 –Low-Creditability Tweets in the US. vs Administered Vaccines in the US.

Table 1– Linear Regression Model of Low-Creditability Tweets in the US vs Administered Vaccines in the US.

Min	**1Q**	**Median**	**3Q**	**Max**
7970459	2954022	599033	650977	11437534

	Estimate Std.	**Error**	**t value**	**Pr(>|t|)**
(Intercept)	8250059	2509053	3.288	0.00387 **
MisInfo	52989	13717	3.863	0.00105 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4576000 on 19 degrees of freedom

Multiple R-squared: 0.4399, Adjusted R-squared: 0.4104

F-statistic: 14.92 on 1 and 19 DF, p-value: 0.001047

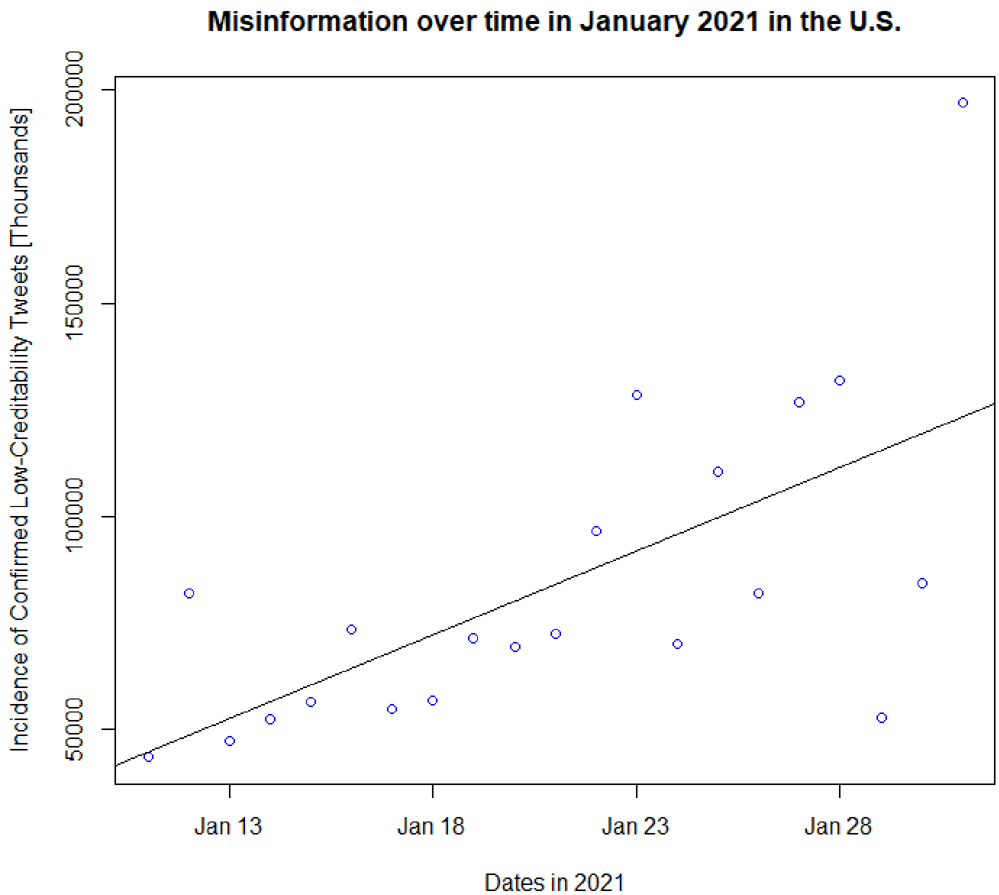

Figure 2 – Low-Creditability Tweets in the US for January 2021.

Table 2 – Multiple Regression Model of Dates and Misinformation on US Administered Vaccines

Min	**1Q**	**Median**	**3Q**	**Max**
1218642	580831	580831	413049	1281756

	Estimate Std.	**Error**	**t value**	**Pr(>|t|)**
(Intercept)	-1.746e+10	6.754e+08	-25.854	1.10e-15 ***
MisInfo	4.064e+00	6.028e+00	0.674	0.509
Dates	1.085e+01	4.194e-01	25.866	1.09e-15 ***

Residual standard error: 761000 on 18 degrees of freedom

Multiple R-squared: 0.9853, Adjusted R-squared: 0.9837

F-statistic: 604.3 on 2 and 18 DF, p-value: < 2.2e-16

Table 3 – Variance Inflation Factor (VIF) for the Multiple Regression Model of Dates and Misinformation on US Administered Vaccines.

MisInfo	Dates
1.745916	1.745916

Discussion

Surprisingly the trend found in Figure 1, as is shown above, does not depict a negative relationship between the quantity of tweets promoting links to articles on low-creditability new sources, herein referred to as Tweet misinformation. Albeit in contrast the trend is positive during this time interval in January of 2021. The P-value < 0.05 with $\alpha = 0.05$ for the coefficients of the linear regression models confirms the significance of X_i (Days in January) relating to Y_i (Administered Vaccines in the US). This means that as the number of misinformation increased more vaccines were taken. However, as the temporal dimension increased so did the quantity of tweet misinformation. Likewise, as the temporal dimension increased so did the number of vaccine administrations. A multiple-regression model was formed using the format followed in equation 1 with β_2 and X_{i2} included in the model of the relation of the Dates and Tweet Misinformation on Vaccine Administration. In this statistical summary the model coefficient for Tweet misinformation was not statistically significant and the model coefficient for Dates was statistically significant (P<0.05 with $\alpha = 0.05$).

This leads to speculation that the independent variables may be correlated. If this is true, data collinearity creates mistrust in integrity of the P-values ability to indicate statically significant independent variables, therefore weakening the model with that respect. The degree of collinearity was assessed using the variance inflation factor (VIF), as a measure of regression coefficient (β_i) variance due to multicollinearity within the model. Table 3 illustrates a VIF value of approximately 1.74 for both terms in the multiple regression model in Table 2. Based upon academic convention for VIF values this indicates that data collinearity between these two variables is not present[30]–[32].

To recapitulate, this confirms the results that the model coefficient for Tweet misinformation is not statistically significant from the multiple regression model with date and number of Tweets containing misinformation as the independent variables and number of US Vaccine Administered as the dependent variable.

There are a number of potential reasons that the quantity of instances of misinformation did not have a relationship with vaccine administration.

(1) The temporal sample duration was too small and was early on in the OWS initiative when popularity of the vaccines was high.

(2) Misinformation quantities were far too low to have an impact on the greater population, however, they were increasing so this may be prior to an inflection point on the vaccine administration rates.

(3) Gaps in the dataset's assignment on tweet misinformation and the assessment of news source creditability. More tweets within the sample from the CoVaxxy database could have contained misinformation than identified using the method described within this study.

(4) Additional sources of misinformation such as other social medias (Facebook, Instagram, Snapchat, etc.) and TV news networks may impact the results on misinformation generating a different effect.

(5) A portion of the population without social media accounts in the population are not affected by this form of misinformation.

(6) A critical assumption that timing of exposure to misinformation from this representative population cross-section experiences uniform exposure.

(7) Improper month selection for sample set based on US Government project OWS.

Conclusion and Recommendations

This analysis suggests that in January of 2021, the tweets published in the U.S. using substrings related to COVID-19 and vaccines that promoted articles to known unreliable new sources did not have a statistically significant impact on vaccine administration in the U.S.

The date sample size of tweets is not sufficient to extend a statement on the overall impact of vaccine related tweet misinformation on U.S. vaccination rates. Focusing on just the social media platform, Twitter, limits the scope of the study as individuals without accounts are not impacted and other sources of information are excluded. It is hypothesized that the vaccination rates naturally plateau as the number of individuals willing to receive the vaccine have already been inoculated, the impact of misinformation on the remaining unvaccinated individuals increases. A study with independent variables such as percentage of population vaccinated and incidence of misinformation on the vaccine administration would be an interesting assessment on this topic. A complimentary study analyzing each region state by state where demands are different may yield different results. Additionally, a system capable of contemporaneously processing and visualizing data on this topic as the situation develops daily would be an extremely useful tool to identify trends in misinformation, vaccine administration rates and other variables.

Acknowledgments

This paper was made possible in part by the invaluable mentorships and guidance on statistical methods provided by Professor Andisheh Dadashi M.S. A special thank you is extended to Jennifer Nieto, Professor, Elaine Clark, and Professor Emeritus, Frank Gilfeather, for their support and guidance within the Critical Technologies Studies Program.

References

[1] H. Ritchie *et al.*, “Coronavirus (COVID-19) Cases - Statistics and Research - Our World in Data,” *Daily new confirmed COVID-19 cases per million people*, May 28, 2021. https://ourworldindata.org/covid-cases (accessed May 28, 2021).

[2] J. Shaman, R. Talbot, A. Carlsen, and N. Aizenman, “Actual Coronavirus Infections Far Higher Than Confirmed Cases, Model Shows : Shots - Health News : NPR,” *CORONAVIRUS BY THE NUMBERS*, Feb. 06, 2021. https://www.npr.org/sections/health-shots/2021/02/06/964527835/why-the-pandemic-is-10-times-worse-than-you-think (accessed May 28, 2021).

[3] Worldometer, “Countries where Coronavirus has spread - Worldometer.” https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/ (accessed May 28, 2021).

[4] W. Yang, E. Petkova, and J. Shaman, “The 1918 influenza pandemic in New York City: Age-specific timing, mortality, and transmission dynamics,” *Influenza Other Respi. Viruses*, vol. 8, no. 2, pp. 177–188, Mar. 2014, doi: 10.1111/irv.12217.

[5] World Health Organization, “Global HIV Programme.” https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/strategic-information/hiv-data-and-statistics (accessed May 28, 2021).

[6] N. P. A. S. Johnson and J. Mueller, “Updating the Accounts: Global Mortality of the 1918-1920 ‘Spanish’ Influenza Pandemic,” 2002. Accessed: May 28, 2021. [Online]. Available: https://about.jstor.org/terms.

[7] “PREDICT - EcoHealth Alliance.” https://www.ecohealthalliance.org/program/predict (accessed May 28, 2021).

[8] N. Madhav, B. Oppenheim, M. Gallivan, P. Mulembakani, E. Rubin, and N. Wolfe, “Pandemics: Risks, Impacts, and Mitigation,” in *Disease Control Priorities, Third Edition (Volume 9): Improving Health and Reducing Poverty*, The World Bank, 2017, pp. 315–345.

[9] V. Hul *et al.*, “A novel SARS-CoV-2 related coronavirus in bats from Cambodia,” *bioRxiv*. Cold Spring Harbor Laboratory, p. 2021.01.26.428212, Jan. 26, 2021, doi: 10.1101/2021.01.26.428212.

[10] E. Ferreri, “Researchers Discuss New Vaccine That Could Prevent Future Pandemics | Duke Today.” https://today.duke.edu/2021/05/researchers-discuss-new-vaccine-could-prevent-future-pandemics (accessed May 28, 2021).

[11] A. J. Wakefield *et al.*, “Retracted: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children,” *Lancet*, vol. 351, no. 9103, pp. 637–641, Feb. 1998, doi: 10.1016/S0140-6736(97)11096-0.

[12] T. S. S. Rao and C. Andrade, “The MMR vaccine and autism: Sensation, refutation, retraction, and fraud,” *Indian Journal of Psychiatry*, vol. 53, no. 2. Wolters Kluwer --

Medknow Publications, pp. 95–96, Apr. 2011, doi: 10.4103/0019-5545.82529.

[13] CDC, “What is Autism Spectrum Disorder? | CDC.” https://www.cdc.gov/ncbddd/autism/facts.html (accessed May 28, 2021).

[14] F. Destefano, C. S. Price, and E. S. Weintraub, “Increasing Exposure to Antibody-Stimulating Proteins and Polysaccharides in Vaccines Is Not Associated with Risk of Autism AD Autistic disorder ADI-R Autism Diagnostic Interview-Revised ADOS Autism Diagnostic Observation Schedule ASD Autism spectrum disorder MCO Managed care organization SCQ Social Communication Questionnaire,” *J. Pediatr.*, vol. 163, pp. 561–567, 2013, doi: 10.1016/j.jpeds.2013.02.001.

[15] “Antigen: MedlinePlus Medical Encyclopedia.” https://medlineplus.gov/ency/article/002224.htm (accessed May 28, 2021).

[16] “Misinformation fueling vaccine hesitancy, PAHO Director says - PAHO/WHO | Pan American Health Organization.” https://www.paho.org/en/news/21-4-2021-misinformation-fueling-vaccine-hesitancy-paho-director-says (accessed May 28, 2021).

[17] U. C. Bureau, “Around Half of Unvaccinated Americans Indicate They Will ‘Definitely’ Get COVID-19 Vaccine,” Accessed: May 28, 2021. [Online]. Available: https://www.census.gov/library/stories/2021/01/around-half-of-unvaccinated-americans-indicate-they-will-definitely-get-covid-19-vaccine.html.

[18] F. Pierri *et al.*, “The impact of online misinformation on U.S. COVID-19 vaccinations.”

[19] V. Carrieri, L. Madio, and F. Principe, “Vaccine hesitancy and (fake) news: Quasi-experimental evidence from Italy,” *Heal. Econ. (United Kingdom)*, vol. 28, no. 11, pp. 1377–1382, Nov. 2019, doi: 10.1002/hec.3937.

[20] M. D. Conover, B. Gonçalves, A. Flammini, and F. Menczer, “Partisan Asymmetries in Online Political Activity,” 2012.

[21] M. R. Deverna *et al.*, “CoVaxxy: A Collection of English-Language Twitter Posts about COVID-19 Vaccines,” 2021, doi: 10.5281/zenodo.4526494.

[22] Twitter, “Twitter API v2,” 2021. https://developer.twitter.com/en/docs/twitter-api/rate-limits (accessed Jun. 21, 2021).

[23] CDC, “CDC COVID Data Tracker,” *COVID-19 Vaccinations in the United States*, Jan. 06, 2021. https://web.archive.org/web/20210106155026/https://covid.cdc.gov/covid-data-tracker/#vaccinations (accessed Jul. 18, 2021).

[24] B. Golding and J. Verbanac, “Iffy+ Mis/Disinfo Sites – Iffy.news.” https://iffy.news/iffy-plus/ (accessed Jul. 18, 2021).

[25] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, “The spread of low-credibility content by social bots,” doi: 10.1038/s41467-018-06930-7.

[26] E. Saltz, C. Leibowicz, and A. C. Wardle, “Encounters with Visual Misinformation and Labels Across Platforms An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions,” 2020.

[27] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, “Fake news on Twitter during the 2016 U.S. presidential election,” Accessed: Jul. 18, 2021. [Online]. Available: http://science.sciencemag.org/.

[28] P. Resnick, A. Ovadya, and G. Gilchrist, “Iffy Quotient: A Platform Health Metric for Misinformation,” Accessed: Jul. 18, 2021. [Online]. Available: http://umsi.info/iffy-quotient-whitepaper.

[29] “COVID-19 Vaccinations in the United States,Jurisdiction | Data | Centers for Disease Control and Prevention.” https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc (accessed Jul. 18, 2021).

[30] “Multicollinearity Essentials and VIF in R - Articles - STHDA.” http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/ (accessed Jul. 22, 2021).

[31] A. Alin, “Multicollinearity,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 3, pp. 370–374, May 2010, doi: 10.1002/WICS.84.

[32] J. H. Kim, “Multicollinearity and misleading statistical results,” *Korean J. Anesthesiol.*, vol. 72, no. 6, p. 558, Dec. 2019, doi: 10.4097/KJA.19087.

Supplemental Instruction

URL Creditability Score

To identify misinformation, Tweets containing URLs are assigned their respective credibility score based on a neutral third-party index known as the IFFY+ Index. Scores from this index are not ascribed at the degree of the story but the degree of the publisher, like the approach taken by other data analyses[25]–[28]. The methodology used to construct the IFFY+ Index of websites that routinely generate or share low-factual information, aligns with the Media Bias / Fact Check (MB/FC) organization's process. These sites regularly violate standards of professional journalism by production misinformation in the form of implicit parodies on news media, fabricated (fake) news, and publisher imposters. It is also unattainable to individually fact check millions of articles therefore the approach of using the publisher's track record provides the most utility. Sites that have a controversial reliability (i.e. fact checkers disagree on the credibility) are excluded from the index, therefore all sites listed are agreed have low-credibility. At the time of this data analysis the IFFY+ Index was composed of 814 different sites.