Aalborg Universitet

"Does Vinegar Kill Coronavirus?" - Using Search Log Analysis to Estimate the Extent of COVID-19-Related Misinformation Searching Behaviour in the United States

Windfeld, Andreas Christian; Meier, Florian Maximilian

Published in:
Proceedings of iConference 2021

Publication date:
2021

Document Version
Early version, also known as pre-print

"Does Vinegar Kill Coronavirus?" - Using Search Log Analysis to Estimate the Extent of COVID-19-Related Misinformation Searching Behaviour in the United States

Andreas Christian Windfeld and Florian Meier

Science, Policy and Information Studies
Department of Communication and Psychology
Aalborg University, Copenhagen, Denmark
awindfeld@gmail.com,fmeier@hum.aau.dk

Abstract. Health experts and government authorities' actions to combat the coronavirus outbreak are strongly compromised by the misinformation infodemic that evolved in parallel to the COVID-19 pandemic. When people get misled by unscientific and unsubstantiated claims regarding the origin or cures for COVID-19, public health response efforts get undermined and people might be less likely to comply with official guidance and thus spread the virus or even harm themselves. To prevent this from happening, a first step is to reveal the prevalence of misinformation ideas in the public. In this study, we use search log analysis to investigate the extent and characteristics of misinformation seeking behaviour in the US using the Bing Search Data-set for Coronavirus Intent. We train a machine learning model to distinguish between regular and misinformation queries and find that only around 1% of queries are related to misinformation myths or conspiracy theories. The query term *qanon* — connecting the conspiracy theory to many different origin myths of COVID-19 — is the most frequent and steadily increasing misinformation-related query in the data-set.

Keywords: COVID-19 · coronavirus · infodemic · misinformation · Bing · search log analysis

1 Introduction

On March 11, 2020, the World Health Organisation (WHO) declared the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes the coronavirus disease COVID-19, a global pandemic [26]. For several months now, health experts and authorities around the globe try to fight the pandemic through developing vaccines, finding novel treatments and communicating what they have learned about the virus to inform people on how to best prevent infection and contain the spread of the virus. However, their work is seriously jeopardized by the infodemic that accompanies the pandemic [28]. An infodemic is defined as "an overabundance of information —— some accurate and some not —— that occurs during an epidemic" [25]. Most problematic about an infodemic is the vast amount of inaccurate, false and misinformation that gets propagated. The spread of misinformation is a serious concern in fighting a pandemic because when people get misled by unscientific and unsubstantiated claims regarding the origin or cures for COVID-19, public health response efforts get undermined and people might be less likely to comply with official guidance and thus spread the virus or even harm themselves by following false claims e.g. injecting disinfectant in seek for protection [1].

In the US, for example, many Republican officials including President Donald Trump, downplayed the severity of the crisis, which lead to less social distancing and more COVID-19 infections in Republican-leaning states [5]. Moreover, a recent study found Donald Trump to be "likely the largest driver of the COVID-19 misinformation infodemic" in English-language news media [12]. Motivated by these observations, we were wondering how widespread the interest in misinformation topics among US citizens is. Thus, in this study, we use search queries as a proxy for public interest and analyse Bing's search logs with COVID-19 intent made publicly available by Microsoft [20] to reveal the extent and characteristics of COVID-19 misinformation-related searching behaviour in the US. We train a classifier to determine whether a query is related to misinformation or not. For training the machine learning model we create a list of COVID-19 misinformation themes and associated keywords based on multiple sources [16,17,12]. We find that only around 1% of the search volume in the data-set is related to misinformation seeking behaviour. While we observe queries related to possible cures (e.g. *hydroxychloroquine coronavirus* or *does vinegar kill coronavirus*), the most frequent misinformation query is associated with the QAnon conspiracy theory propagating various origin myths of the virus.

2 Related Work

Mining query logs for studying public interest in health topics has a long tradition in health information behavior research [8,11,1]. Our study is mostly situated in the context of *supply-based infodemiology* which assesses the quality of online health (mis-)information [13], tries to predict epidemic outbreaks from search log data [15] and investigates the public's reaction to epidemics [7].

Although COVID-19 is only several month old, many studies already discuss or empirically study the infodemic following it [3,21,17,19,24,28]. Rovetta et al. studied web search behavior and infodemic attitudes in Italy using Google trends [24]. They find that misinformation (e.g. 5G coronavirus) was widely circulated in the Campania region and racism-related information (e.g. chinese virus) in Umbria and Basilicata. Islam et al. classify COVID-19 misinformation into rumours, stigma, and conspiracy theories and study the volume of misinformation using social media data [17]. Their main finding is that misinformation

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters#bleach

Country	# Queries	% Queries
United States	1751769	45.73
United Kingdom	805201	21.02
France	197275	5.15
Italy	172992	4.52
Germany	166362	4.34
Canada	144528	3.77
Japan	78967	2.06
Spain	73105	1.91
Australia	71794	1.87
India	49231	1.29

Table 1: Count and share of queries per top ten countries in the data-set.

is most prominent in USA and India and that misinformation has the potential to decrease community trust in governments and international health agencies. Makhortykh, Urman, and Ulloa perform an analysis of search engine results for "coronavirus" in English, Russian, and Mandarin highlighting the significant differences in the types of resources provided to users across search engines and languages [19].

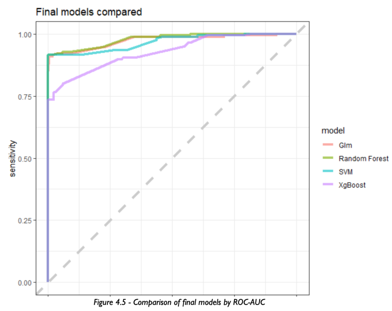
3 Data-set & Experiments

The *Bing Search Data-set for Coronavirus Intent* is a curated set of Bing search logs published by Microsoft with new updates released on Github every month [20]. The data-set contains queries issued by desktop users with intent related to the coronavirus or COVID-19. In some cases, this intent is explicit in the query itself e.g., *Coronavirus map*, in other cases it is implicit, e.g., *toilet paper*. The implicit intent is predicted by a method known as random walks on the click graph [10]. Instead of raw query frequencies, the popularity of a query is represented in a normalized *popularity score* ranging from 1 - 100.

The version of the data-set we use in our study covers the time period from January 1st to August 31st 2020 and contains 3,830,284 queries coming from all over the world. Table 1 lists the ten countries with the largest volume in the data-set and shows that queries issued in the US account for almost half of all queries (45.73%) while other countries are only poorly represented. The top five most frequent queries can be seen in Figure 3.

3.1 Building a Classification Model

To identify misinformation queries in the data-set we trained a classifier to distinguish between regular and misinformation-related queries. The model was trained in a three-step approach: (1) Simply selecting a random sample of queries



(a)

Model	Sensitivity	Specificity	ROC-AUC
Glm	0.910	0.991	0.974
Random Forest	0.916	1	0.979
XGBoost	0.452	1	0.928
SVM	0.916	1	0.967

(b)

Fig. 1: ROC-AUC (a) visualiszation and (b) values for the final models.

for manual coding to create a ground truth data-set for training was deemed unfeasible as initial analysis showed that misinformation queries are simply not frequent enough to create a balanced training data-set. Thus we created a list of misinformation themes and associated keywords relating to conspiracy theories and myths around COVID-19. To create this list we consulted different sources e.g. The WHO Mythbusters resource[2], the COVID19MisInfo.org[3] portal [16] and similar lists created by Islam et al. [17] and Evanega et al. [12]. (2) We used this list to create keyword-based regular expressions that helped with randomly selecting 1000 queries and building the positive class. We supplemented this 1000 queries with another random sample of 1000 queries representing regular queries i.e. the negative class. The 2000 queries were manually coded by two researchers on whether a query is with regular or misinformation intent. A total of 52 observations were found to be coded differently (Cohen's Kappa $\kappa = 0.946$). The data was split into a training and test set at an 80/20 ratio. (3) The supervised learning part started with feature engineering. We created *idf* feature vectors based on unigrams and bigrams. We experimented with selecting different numbers of most frequent tokens (20-200). We compared four algorithms: Logistic regression (glmnet,[14]), random forest (ranger,[27]), tree boosting (XGBoost,[9]) and support vector machines (SVM). Hyperparameters were tuned when possible. All machine learning was performed using 10-fold cross-validation with splits set to be stratified according to the original partition. As we were mainly interested in the extent of misinformation related searching behaviour we optimized the model for sensitivity (recall).

Figure 1a and Figure 1b show the results of the final and tuned models using ROC-AUC. The performance was excellent across all models excluding XGBoost, which had a very low sensitivity score of 45%. Glm, random forest, and SVM all had sensitivity scores above 90% and ROC-AUC scores above 96%.

[2] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters
[3] https://covid19misinfo.org/

	Reg. Total	Misinfo. Total	% Misinfo	Reg. Unique	Misinfo. Unique	% Reg. Unique	%. Misinfo Unique
Jan	16.542	796	4.8%	1.090	21	6.6 %	2.6%
Feb	50.426	654	1.3%	2.346	34	4.7%	5.5%
Mar	430.810	5.217	1.2%	22.544	143	5.2%	2.7%
Apr	380.091	4.312	1.1%	14.446	95	3.8%	2.2%
May	213.785	1.597	0.7%	8.129	54	3.8%	3.4%
Jun	212.325	1.187	0.6%	7.522	36	3.5%	3%
Jul	249.758	1.878	0.8%	8.648	44	3.5%	2.4%
Aug	180.744	1.647	0.9%	5.570	21	3.1%	1.3%

Fig. 2: Share of regular and misinformation queries per month.

Given the very similar performance between the models, the logistic regression model (*glm*) was selected as it is the simpler model with lower computational requirements [2].

4 Results

Applying the final model to the whole data-set revealed only around 1% (0.98%; 17.288) of all queries being classified as misinformation-related searches. This is quite low considering the fact that recent research found "over 1.1 million news articles [...] that disseminated, amplified or reported misinformation related to the pandemic" [12, p.3]. Considering this high number of news reports containing misinformation one might expect to see more search volume dedicated to it.

4.1 National-Level Analysis

Figure 2 gives an overview of the number of regular and misinformation queries from January to August. One can see that the search volume is highest in March (430,810) and April (380,091), when in most countries the first wave of infections was hitting. Moreover, it shows that the share of misinformation-related queries is highest in January (4.8%) and balances out to be around 1% in the following month. Table 2 also lists the number and share of unique regular and misinformation queries in this time period. While the share of unique regular queries varies between 3.1% and 6.6%, the share of unique misinformation-related queries ranges only from 1.3% to 5.5%. In total, we observed only 378 unique misinformation queries. This, suggests that misinformation searching behaviour is focused on a smaller set of themes or topics. Regular queries are also more popular with an average popularity score of 4.85 (misinformation queries popularity score $avg = 2.62$).

Figure 3 visualizes the ten most frequent regular and misinformation-related queries side-by-side and reveals *qanon* to be the most frequent misinformation search term, occurring 4685 times within the data-set period. *Qanon* queries are labelled as *implicit intent*, meaning their relation to the coronavirus pandemic

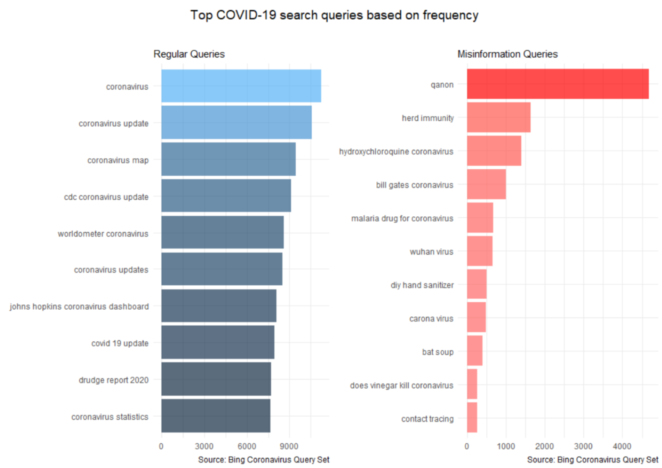


Fig. 3: Most frequent regular and misinformation queries.

is based on the users clickstream-interaction. The QAnon conspiracy theory is originally revolving around US President Donald Trump fighting a satanic cult of elite-paedophiles forming a secret government (deep state). However, supporters of QAnon started to also spread misinformation regarding the COVID-19 pandemic e.g. the pandemic being a population control scheme [17] or the technology standard 5G causing the infection [3].

The query *herd immunity* refers to the idea of a "large uncontrolled outbreak in the low-risk population while protecting the vulnerable. Proponents suggest this would lead to the development of infection-acquired population immunity in the low-risk population" [4]. However, most scientists and healthcare professionals consider this strategy to be a "dangerous fallacy unsupported by scientific evidence" [4] that would cause many deaths and suffering, "but not speeding up society's return to normal" [6]. Other queries relate to myths around cures (*hydroxychloroquine coronavirus*, *malaria drug coronavirus*, *does vinegar kill coronavirus*) or the origin of the virus (*bat soup*). The misspelled query *carona virus* is a false positive.

Finally, we looked at the development of the five most frequent misinformation queries (seen in Figure 3) over time. Figure 4 visualizes their development and shows that neither of them was present in January. In February, we begin to see the first occurrences. It is interesting to note, that the query *hydroxychloroquine coronavirus* sparked in April, after President Donald Trump mentioned the antimalarial drug as a treatment for COVID-19 at a press briefing on March 19 [22]. Moreover, we can see the myth around Bill Gates' involvement in the spread of the virus loses popularity. From June onward no instances of the query *bill gates coronavirus* can be seen. Most striking is the development of *qanon* which increases steadily and peaks in August. The increase is also reflected in *qanon*'s popularity score which surged from 1 in January, February and March

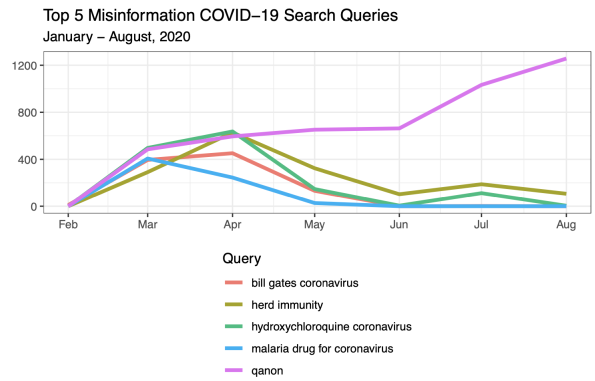


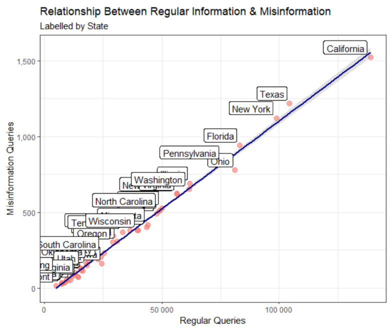
Fig. 4: Time series analysis of the top five misinformation queries.

to 60 in August. As the conspiracy theory related movement is gaining more and more popularity in the US, their views on the coronavirus outbreak attracts more and more attention.

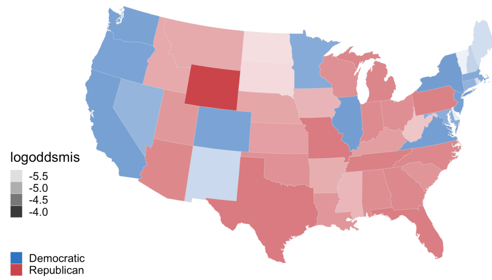
4.2 Per-state-Level Analysis

US states showed vast differences in how they were affected by the spread of COVID-19. While some government officials downplayed the virus, others enforced interventions for stopping the spread like mandates to wear face mask and practice social distancing [23]. This motivated us to also investigate per-state-level differences regarding the extent of how much people engage in misinformation-related searching behaviour. Figure 5a visualizes the number of regular queries in relation to the number of misinformation queries. One can observe that the US states with the highest population (California, Texas and Florida) also do have the most search volume. Moreover, the regression line that was produced by a linear model is diagonal with almost all states aligning perfectly along with it. This suggests that no state had remarkably more or less misinformation-related searches.

Studies found political orientation and partisanship to be strongly associated with the spread of the coronavirus as people in pro-Trump states have tended to practice fewer social distancing behaviors and displayed less concern over COVID-19, resulting in more infections [5,18]. Thus we investigated whether the political orientation of a state would lead to more misinformation searching behaviour. Figure 5b shows a map of the US. States are colored by being either Republican-leaning or Democratic-leaning based on the 2016 election outcome. The color saturation of a state represents the likelihood of a misinformation query being observed (log-odds). The more saturated a state appears the more likely it is that their residents engage in misinformation searching behaviour. Figure 5b reveals that the likelihood of observing a misinformation related query is highest in the three Republican-leaning states Wyoming (-3.97), Missouri (-4.45) and



(a)



(b)

Fig. 5: (a) Relation between the number of regular and misinformation queries (counts) and (b) visualisation of political orientation (election result 2016) and likelihood of misinformation-related queries (log-odds) per state.

Texas (-4.45). However, the Democratic-leaning state New York follows closely with $logodds = -4.47$. Among the top ten states, five are Republican-leaning and five are democratic-leaning. Thus, in general, we don't observe a tendency or evidence for political affiliation explaining the volume of misinformation searches.

5 Discussion and Conclusion

This study performed a search log analysis of the Bing search data-set for Coronavirus intent using a supervised machine learning model to classify queries into regular and misinformation-related queries with the aim to reveal the extent and characteristics of misinformation searching behaviour in the US.

The overall extent of misinformation, was observed to be approximately 1%. Given the huge volume of media articles covering the COVID-19 infodemic [12] the interest in misinformation themes appears to be rather low. Of course it is questionable whether the data is representative enough as search queries are pre-selected by Microsoft and the market share of Bing only lies at 2.83%[4]. We identified the level of misinformation being highest in January (4.8%), but several top misinformation queries of this month included misspellings of *coronavirus* which can be considered false positives. These need to be revisited in future analysis.

The most frequent query by far was *qanon*. Other queries are related to myths about the origin (*bat soup*) and treatment (*hydroxychloroquine coronavirus*) of the coronavirus, some are related to popular conspiracy theories (*bill gates coronavirus*), and few queries could be interpreted as assigning blame or stigma towards Asian groups (*wuhan virus*). An analysis of whether political orientation (section 4.2) has an effect on the extent of misinformation queries did not reveal

[4] https://gs.statcounter.com/search-engine-market-share

any evidence for this claim. However, the Republican-leaning state Wyoming was identified as the state with the highest likelihood for misinformation queries.

The study showed, that conspiracy theories and myths that prompt misinformation searching behaviour are bursty in nature often changing rapidly. While some myths — like the malaria drug hydroxychloroquine being an effective cure — rose and disappeared during the data-set period, *qanon* rose steadily.

References

1. Abebe, R., Hill, S., Vaughan, J.W., Small, P.M., Schwartz, H.A.: Using search queries to understand health information needs in africa. Proceedings of the International AAAI Conference on Web and Social Media **13**(01), 3–14 (2019), https://aaai.org/ojs/index.php/ICWSM/article/view/3360
2. Aggarwal, C.C.: Classifier Performance and Evaluation, pp. 209–234. Springer International Publishing, Cham (2018), https://doi.org/10.1007/978-3-319-73531-3_7
3. Ahmed, W., Vidal-Alaball, J., Downing, J., López Seguí, F.: Covid-19 and the 5g conspiracy theory: Social network analysis of twitter data. J Med Internet Res **22**(5) (2020), http://www.jmir.org/2020/5/e19458/
4. et al., N.A.A.: Scientific consensus on the covid-19 pandemic: we need to act now. Lancet (2020), https://doi.org/10.1016/S0140-6736(20)32153-X
5. Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., Yang, D.: Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. Journal of Public Economics **191**, 104254 (2020). https://doi.org/https://doi.org/10.1016/j.jpubeco.2020.104254, http://www.sciencedirect.com/science/article/pii/S0047272720301183
6. Aschwanden, C.: The false promise of herd immunity for covid-19 (2020), https://doi.org/10.1038/d41586-020-02948-4
7. Bragazzi, N.L., Alicino, C., Trucchi, C., Paganino, C., Barberis, I., Martini, M., Sticchi, L., Trinka, E., Brigo, F., Ansaldi, F., Icardi, G., Orsi, A.: Global reaction to the recent outbreaks of zika virus: Insights from a big data analysis. PLOS ONE **12**, 1–15 (09 2017), https://doi.org/10.1371/journal.pone.0185263
8. Cartright, M.A., White, R.W., Horvitz, E.: Intentions and attention in exploratory health search. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 65–74. SIGIR '11, Association for Computing Machinery, New York, NY, USA (2011), https://doi.org/10.1145/2009916.2009929
9. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. KDD '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2939672.2939785
10. Craswell, N., Szummer, M.: Random walks on the click graph. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 239–246. Association for Computing Machinery, New York, NY, USA (2007), https://doi.org/10.1145/1277741.1277784
11. De Choudhury, M., Morris, M.R., White, R.W.: Seeking and sharing health information online: Comparing search engines and social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p. 1365–1376. CHI '14, Association for Computing Machinery, New York,

NY, USA (2014). https://doi.org/10.1145/2556288.2557214, https://doi.org/10.1145/2556288.2557214
12. Evanega, S., Lynas, M., Adams, J., Smolenyak, K.: Coronavirus misinformation:quantifying sources and themes in the covid-19 'infodemic' (2020), https://allianceforscience.cornell.edu/wp-content/uploads/2020/09/Evanega-et-al-Coronavirus-misinformationFINAL.pdf
13. Eysenbach, G.: Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res **11**(1), e11 (2009), https://doi.org/10.2196/jmir.1157
14. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**(1), 1–22 (2010), http://www.jstatsoft.org/v33/i01/
15. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature **457**(7232), 1012–1014 (2009), https://doi.org/10.1038/nature07634
16. Gruzd, A. and Mai, P.: COVID-19 Misinformation Portal – A Rapid Response Project from the Ryerson University Social Media Lab. https://covid19misinfo.org/ (2020), accessed: 2020-10-25
17. Islam, M.S., Sarkar, T., Khan, S.H., Mostofa Kamal, A.H., Hasan, S.M.M., Kabir, A., Yeasmin, D., Islam, M.A., Amin Chowdhury, K.I., Anwar, K.S., Chughtai, A.A., Seale, H.: Covid-19–related infodemic and its impact on public health: A global social media analysis. The American Journal of Tropical Medicine and Hygiene **103**(4), 1621–1629 (2020), http://www.ajtmh.org/content/journals/10.4269/ajtmh.20-0812
18. Kubinec, R., Carvalho, L., Barceló, J., Cheng, C., Messerschmidt, L., Duba, D., Cottrell, M.S.: Fear, partisanship and the spread of covid-19 in the united states (2020), osf.io/preprints/socarxiv/jp4wk
19. Makhortykh, M., Urman, A., Ulloa, R.: How search engines disseminate information about covid-19 and why they should do better. The Harvard Kennedy School (HKS) Misinformation Review - Special Issue on COVID-19 and Misinformation **1** (2020), http://publichealth.jmir.org/2020/2/e19374/
20. Microsoft: Data source: Bing coronavirus query set (2020), https://github.com/microsoft/BingCoronavirusQuerySet
21. Orso, D., Federici, N., Copetti, R., Vetrugno, L., Bove, T.: Infodemic and the spread of fake news in the COVID-19-era. European Journal of Emergency Medicine (2020), https://bit.ly/3giWRXt
22. Rachel Sandler: Hydroxychloroquine Abuse Up Since Trump First Mentioned Drug, U.S. Poison Center Data Says. https://www.forbes.com/sites/rachelsandler/2020/04/09/hydroxychloroquine-abuse-up-since-trump-first-mentioned-drug-us-poison-centers-say/ (2020), accessed: 2020-10-25
23. Reiner, R., Barber, R., Collins, J.e.a.: Modeling covid-19 scenarios for the united states. Nature Medicine (2020), https://doi.org/10.1038/s41591-020-1132-9
24. Rovetta, A., Bhagavathula, A.S.: Covid-19-related web search behaviors and infodemic attitudes in italy: Infodemiological study. JMIR Public Health Surveill **6**(2), e19374 (May 2020), http://publichealth.jmir.org/2020/2/e19374/
25. Tangcharoensathien, V., Calleja, N., Nguyen, T., Purnat, T., D'Agostino, M., Garcia-Saiso, S., Landry, M., Rashidian, A., Hamilton, C., AbdAllah, A., Ghiga, I., Hill, A., Hougendobler, D., van Andel, J., Nunn, M., Brooks, I., Sacco, P.L.,

De Domenico, M., Mai, P., Gruzd, A., Alaphilippe, A., Briand, S.: Framework for managing the covid-19 infodemic: Methods and results of an online, crowd-sourced who technical consultation. J Med Internet Res **22**(6), e19659 (2020), http://www.jmir.org/2020/6/e19659/
26. Tedros Adhanom Ghebreyesus: WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020 (2020), accessed: 2020-10-25
27. Wright, M.N., Ziegler, A.: ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software **77**(1), 1–17 (2017). https://doi.org/10.18637/jss.v077.i01
28. Zarocostas, J.: How to figh an infodemic? The Lancet **395** (2020), https://doi.org/10.1016/S0140-6736(20)30461-X