A Google–Wikipedia–Twitter Model as a Leading Indicator of the Numbers of Coronavirus Deaths

Daniel E. O'Leary1 | Veda C. Storey2

1University of Southern California, Los Angeles, CA, USA
2Georgia State University, Atlanta, GA, USA

Correspondence
Daniel E. O'Leary, University of Southern California, Los Angeles, CA, USA.
Email: oleary@usc.edu.

Summary
Forecasting the number of cases and the number of deaths in a pandemic provides critical information to governments and health officials, as seen in the management of the coronavirus outbreak. But things change. Thus, there is a constant search for real-time and leading indicator variables that can provide insights into disease propagation models. Researchers have found that information about social media and search engine use can provide insights into the diffusion of flu and other diseases. Consistent with this finding, we found that a model with the number of Google searches, Twitter tweets, and Wikipedia page views provides a leading indicator model of the number of people in the USA who will become infected and die from the coronavirus. Although we focus on the current coronavirus pandemic, other recent viruses have threatened pandemics (e.g. severe acute respiratory syndrome). Since future and existing diseases are likely to follow a similar search for information, our insights may prove fruitful in dealing with the coronavirus and other such diseases, particularly in the early phases of the disease.

Subject terms: coronavirus, COVID-19, unintentional crowd, Google searches, Wikipedia page views, Twitter tweets, models of disease diffusion.

1 | INTRODUCTION

The coronavirus disease 2019 (COVID-19) is a newly discovered disease, causing respiratory illness. The disease was first identified in December 2019, in Wuhan, China, and has since spread globally, resulting in a pandemic (World Health Organization, 2020). This paper provides an analysis of data early in the coronavirus life cycle, with the data-gathering process ending on 5 April 2020. Although the pandemic is global, our focus is primarily on the USA and English language sources of information because, at the time of writing, the USA had just become the world leader in the number of active coronavirus cases.

Although researchers have noted the importance of many aspects of the pandemic (e.g. O'Leary, 2020), forecasting the number of cases and deaths is, perhaps, one of the most important issues. Unfortunately, researchers have noted that it has been difficult to forecast cases of the coronavirus (Cohen, 2020). Further, Butler (2013) and others have noted the extent to which disease forecasting models need to change over time. In addition, some researchers expect more pandemics in the future. As noted by Haseltine (2020) 'Once this new epidemic has faded into memory, will it be five or 10 years until the next one?' As a result, it is important to continue to search for, and better understand, variables that provide a leading indicator of the number of coronavirus cases and deaths in a range of different settings. As noted by Milinovich, Williams, Clements, and Hu (2014), we need surveillance systems to constantly monitor for infectious diseases. This paper does so by building a model using data from two types of social media (Wikipedia and Twitter tweets) and Google searches to estimate the number of coronavirus cases and deaths in the USA early in the life cycle of the disease. Althouse et al. (2015) referred to these data, such as Google searches, Twitter posts, Wikipedia access, and other data streams as 'novel data streams' (NDSs) and suggested the use of NDSs in models of public health.
1.1 | Gathering Data from the 'Unintentional' Crowd

While visiting Russia in the 1990s, one of the authors rapidly discovered that if people formed a line there was often something considered valuable at the end of it. One line on the street led to the opportunity to buy ice cream; another to buy bread. Thus, the existence of a 'crowd' in line provided data indicating that there was something 'good' or worthwhile at the base of the line. As a more recent example, Fleet (2020) argued that Shanghai is recovering from the coronavirus. He suggested that the existence of a line at a local stuffed bun shop ('baozi shop') was an indicator of that recovery. In each case, the crowd was 'unintentionally' creating information. The people in the line did not intend to draw attention, but by virtue of being in line they were signalling importance (or value) to others.

In a digital world, the same phenomena occur, but there is no physical line. Instead, there may be a user log, or a count of users, or a count of what people asked for or did. We discuss three digital examples: Google, which provides Google Trends, a normalized count of the number of daily internet searches for a word; Wikipedia, which provides a count of the number of daily views for each of its pages; and Twitter, which provides a social media exchange environment.

In the case of disease search or discussion information, the 'unintentional' crowd provides tweets and page view information as the crowd seeks or communicates about information on diseases because they experience symptoms, know others who do, or are curious. The first two reasons suggest that these searches and communications could provide useful insights into the number of potential cases that might occur. In the same sense that those in a physical line become part of an unintentional crowd, Google searches, Twitter tweets, and Wikipedia page views provide the same kind of information. However, that information is digital and usable by anyone who accesses the statistics. Thus, this paper investigates these digital signals to determine their relationship to the coronavirus pandemic, specifically for their ability to forecast the number of cases and/or deaths.

1.2 | Purpose of this Research

Government officials, healthcare managers, and the general public have concerns associated with the coronavirus beyond how to simply stop its spread. Government and healthcare planners focus on having enough resources to handle the number of coronavirus cases and deaths. For example, planners would use the expected number of cases to plan the need for the number of beds and respirators and ventilators. Reportedly, many citizens apparently do not trust the information being provided by the US Government (Montanaro, 2020). As a result, since there is open information, the general public can build their own coronavirus diffusion models. Similarly, researchers need to build models, which requires them to identify what variables are generally related to the coronavirus and which ones, specifically, can help predict the number of cases and deaths.

Our objectives are as follows: to investigate the extent to which Google searches, Twitter tweets, and Wikipedia page views might generate a model that provides a leading indicator of both the number of cases and the number of deaths related to the coronavirus; and to provide the implications for future work on the development of models of disease diffusion. Ultimately, we build a statistical model of the number of coronavirus cases and deaths from which we conclude that the best model includes information from all of these three sources. In addition, we build a model of the number of cases and deaths by state and find that Google searches are a leading indicator.

2 | PREVIOUS LITERATURE: GOOGLE SEARCHES, TWITTER TWEETS, AND WIKIPEDIA PAGE VIEWS AND PUBLIC HEALTH

In a pioneering paper, Ginsberg et al. (2009) suggested that Internet use data, such as Google searches, could be used to monitor diseases, such as the flu. As noted by Ginsberg et al. (2009, p. 1), ‘One way to improve early detection is to monitor health-seeking behaviour in the form of online web search queries, which are submitted by millions of users around the world each day’. Google flu efforts were aimed at analysing searches that had symptoms of the flu. Unfortunately, those efforts have been criticized. Butler (2013) provided a number of arguments, including the strength of existing models and the need to constantly adapt the models. However, Lazer, Kennedy, King, and Vespignani (2014), Kugler (2016) and Nuti et al., (2014) identified methodological concerns associated with the analysis from the perspective of big data and noted the symptoms of the flu and cold were similar, limiting the model. Although the Google flu trends team terminated their efforts in 2015 (Google, 2015), there is continued research using Google searches, as a way of investigating disease outbreaks (e.g. Butler, 2013; Lazer et al., 2014; Kugler, 2016).

Researchers also investigated other social media for their ability to forecast diseases. Ritterman, Osborne, and Klein (2009) used the number of Twitter tweets to monitor the potential for flu and other diseases. Paul and Dredze (2011) also investigated how Twitter could be analysed to forecast diseases. O'Leary (2015) provides a recent summary of ‘Twitter mining’ research.

Tausczik, Faasse, Pennebaker, and Petrie (2012) investigated both the language used in web visits and the number of Wikipedia visits to trace the public’s reaction to the 2009 H1N1 outbreak. McVern and Brownstein (2014) suggested that Wikipedia article views be used as a real-time estimate of the level of influenza-like illness. Priedhorsky et al. (2017) also examined forecasting disease occurrences using Wikipedia information and summarized some of the findings across several diseases, including flu and dengue fever. Aramaki, Maskawa, and Morita (2011) used Twitter to forecast the flu.

Although these efforts were largely oriented at the ability of single media to forecast influenza and other diseases, some researchers have begun to include multiple media in their models. Bardak and Tan (2015) integrated Google Trend data and Wikipedia to predict influenza outbreaks, generating better forecasts than either alone; however, their analysis did not include Twitter data. Sharpe et al. (2016)
compared each of Google Trends, Twitter, and Wikipedia by examining which corresponded best with Centers for Disease Control and Prevention flu data. However, they did not build a model that included all three NDSs. Unfortunately, in our literature review, we did not find the use of Google searches, Twitter tweets, and Wikipedia uses in the estimation of the flu. Further, at the time we gathered our data there was limited publicly available information about forecasting models for the coronavirus.

2.1 Our Approach

In our analysis, we coupled data from three different novel data streams to estimate both the number of coronavirus cases and the number of deaths. As with human experts (e.g., O’Leary 1998) we expect multiple data sources to be better than a single data source. We provide a number of findings. We found, empirically different timing associated with each media, when used together. First, Google Trends was the best at forecasting the coronavirus from 2–3 weeks ahead. Second, Twitter Tweets provided a strong intermediary estimate at 1–2 weeks ahead. Finally, Wikipedia provided a strong estimate up to 7 days ahead closer to the actual events. We also found empirically better results when the number of coronavirus cases was lagged 7 days behind the number of deaths.

3 DATA, VARIABLES, AND METHODOLOGY

Rather than using symptoms of the flu or coronavirus, this research pursues an analysis of direct searches of the name ‘coronavirus’. We do so because the disease was given substantial publicity with its advent in China in December 2019 and January 2020. Before the onset of the coronavirus, people knew only symptoms of the flu, and were unaware of symptoms of the coronavirus. Thus, it is reasonable to expect people to directly search the disease name if they were concerned about the disease.

On 5 April 2020, we extracted data from Google Trend searches and Wikipedia page views for ‘coronavirus’ and ‘COVID-19’. Wikipedia pages for coronavirus started on 12 January 2020, and Google searches were ‘<1’ on Google Trends. The first measurable Google searches apparently started on 21 January 2020. Wikipedia pages and Google searches for COVID-19 started on 11 February 2020 with the advent of the virus’s new name. Despite the new name, the number of page views and searches for ‘coronavirus’ were affected only marginally. For example, on 22 March 2020, the number of Wikipedia page views was approximately 400,000 for coronavirus and 20,000 for COVID-19. As a result, our primary focus has been on ‘coronavirus’.

We also obtained a time series of the number of Twitter tweets, starting on 27 January 2020 from Banda et al. (2020). This set of over 150 million tweets related to the coronavirus pandemic were released on 5 April 2020. We used the daily number of ‘clean’ tweets, with no retweets. Finally, we extracted the numbers of cases and deaths in the USA from www.worldometers.info.

We gathered data pertaining to the six variables described in Table 1.

We used correlation and regression analysis to study the relationships between the variables selected, and the variable inflation factors to measure and limit multicollinearity between the variables in our regression models (Pan & Jackson, 2008). Throughout, we used SAS’s JMP.

4 FINDINGS

We have four primary findings. First, as of the writing of this paper, although the disease has a formal name of COVID-19, the primary searches and page views have been for ‘coronavirus’. Second, by analysing aggregated data from all the USA, we built a model that is a leading indicator of both the number of cases and deaths using each of Wikipedia page views, Google searches, and Twitter tweets. Third, we built a model of the number of cases and deaths by state and find that Google searches provide a leading indicator. Fourth, we find that, to date, there have apparently been three ‘waves’ of searches, tweets, and page views, which suggests that users are engaged for different reasons at different times and that there will be other waves of searches and messages in the future.

4.1 What Is in a Name? COVID-19 versus Coronavirus

On 11 February 2020, what had been known as the ‘coronavirus’ was officially labelled ‘COVID-19.’ The Wikipedia page was created on the same day, and for the first and second days the number of

| Variable                                      | Description                                      |
|-----------------------------------------------|--------------------------------------------------|
| Wikipedia (coronavirus) lag                   | Coronavirus page views in Wikipedia, lag in days  |
| Google (coronavirus) lag                      | Relative number of Google searches for coronavirus, lag in days |
| Twitter lag                                   | Number of Twitter tweet messages—cleaned list, lag in days |
| Google-date                                   | Total number of Google searches by date (GoogleTrends) |
| Daily deaths                                  | Number of daily deaths from the coronavirus      |
| Daily cases                                   | Number of daily cases of the coronavirus         |
page views of COVID-19 compared with coronavirus was 8.06% and 16.66% respectively. After that, the maximum percentage was 8.66%, with the majority being between 4 and 5%. In terms of Google searches, as of 24 March 2020, the ratio of COVID-19 to coronavirus was less than 5%. As a result, although we investigated both, our primary focus was on ‘coronavirus’ in terms of Wikipedia and Google searches.

### 4.2 Models of the Number of Cases and Deaths in the USA

We anticipate that the reporting of the number of coronavirus cases initially was flawed and ‘lumpy’, because of a lack of coronavirus awareness and limited coronavirus testing capabilities. That is apparent in the data, where, reportedly, there were 12 cases on 15 February but none until 22 February, when there were 19 reported. In an effort to use publicly available data that is as accurate as possible, we used data from 25 February through 4 April 2020, a total of 40 observations, to study both the number of cases and deaths.

Previous research has found that the Wikipedia page views are a ‘real-time’ estimator of disease (McIver & Brownstein, 2014), whereas Google searches (Ginsberg et al., 2009) and Twitter tweets (Ritterman et al., 2009) have been predictive. This suggests that the Wikipedia page views are not as much of a leading indicator as the Google searches or Twitter tweets. In addition, because of the nature of the virus, if the coronavirus results in death there is likely to be a week or two between the diagnosis and the death. This suggests that the variables in the model of deaths should be further lagged, by a week or two, more than those in the models of cases. Finally, in order to minimize multicollinearity in our models, we found it important to use the data for each of the variables from different weeks.

Accordingly, in the model of the number of cases we used current Wikipedia page views, 1-week-ahead tweets, and 2-week ahead Google searches. For the model of the number of deaths, we put each of those variables 1 week further ahead. The results for estimating the number of cases and the number of deaths are summarized in Table 2.

The $R^2$ in each of the two models exceeds 0.900, which suggests a strong leading indicator model. Each of the coefficients on the Wikipedia page views, Twitter tweets, and the Google searches in the six models was statistically significant. Further, the signs on coefficients on the Wikipedia page views, Twitter tweets, and Google searches are the same for both the number of cases and the number of deaths. While doing our analysis, we found better models with the Google searches lagged further back than the Twitter tweets.

| Table 2 | Estimate of the number of coronavirus cases and deaths |
|---------|-----------------------------------------------|
|         | Deaths                                      | Cases                                      |
| $R^2$   | 0.9209                                      | 0.9093                                     |
| Adjusted $R^2$ | 0.9143                                      | 0.9017                                     |
| $N$     | 40                                           | 40                                         |
| Constant | −70.664                                      | 2367.1                                     |
| $p$-value | 0.1085                                      | 0.2068                                     |
| Wikipedia (coronavirus) | −0.0208                                     | <0.0001                                    |
| $p$-value |                                         |                                             |
| Wikipedia (coronavirus) – 7 | −0.00041                                    |                                             |
| $p$-value | 0.0061                                      |                                             |
| Tweet – 7   | 13.404                                      | 0.0005                                     |
| $p$-value |                                         |                                             |
| Tweet – 14   | 0.532                                       | 0.0002                                     |
| $p$-value |                                         |                                             |
| Google (coronavirus) – 14 | 138.622                                     | 0.0001                                     |
| $p$-value |                                         |                                             |
| Google (coronavirus) – 21 | 7.985                                       | <0.0001                                    |
| $p$-value |                                         |                                             |
multivariate with robust estimates, and Kth nearest neighbour (K = 8), approaches each identified ‘New York’ State as the only outlier. To limit the possibility that one observation would drive the results, we thus eliminated New York from our data set. Further testing, though, revealed that similar results occurred when we included New York.

The results are summarized in Tables 3 and 4. The results suggest that the number of Google searches, in each of the states, provides a leading indicator of the number of cases and the number of deaths, 2–3 weeks before they occur, consistent with the progression of the coronavirus disease.

Neither Wikipedia nor the Twitter tweets were available on a state-by-state basis; therefore, we could not directly replicate the analysis for tweets or page views.

4.4 Coronavirus Google and Wikipedia: Normal-3 Mixture Distributions

We found that Google searches, Twitter tweets, and Wikipedia views apparently come in ‘waves.’ Using JMP, we choose the distribution that best fits the data for the coronavirus Google searches, Twitter tweets, and Wikipedia page views. Two of the graphs are given in Figures 1 and 2. In each case, a normal-3 mixture was chosen. Such a distribution is consistent with sequential normal distributions occurring over time as the crowd processed information about the coronavirus.

### TABLE 3 Models of the number of cases on 24 March 2020 using Google searches by state

| Date of Google | 3 March | 10 March | 17 March | 20 March | 24 March |
|---------------|---------|----------|----------|----------|----------|
| $R^2$         | 0.0462  | 0.1066   | 0.1072   | 0.1109   | 0.1002   |
| Adjusted $R^2$| 0.0264  | 0.0888   | 0.0886   | 0.0924   | 0.0815   |
| Intercept     | −53.37  | −540.73  | −1498.52 | −0.195   | −1826.04 |
| p-value       | 0.9097  | 0.2905   | 0.1014   | 0.0729   | 0.9049   |
| Google 3 March| 13.11   |          |          |          |          |
| p-value       | 0.1335  |          |          |          |          |
| Google 10 March| 20.19  |          |          |          |          |
| p-value       | 0.0207  |          |          |          |          |
| Google 17 March|        | 27.24   |          |          |          |
| p-value       | 0.0203  |          |          |          |          |
| Google 20 March|        |          |          | 31.88    |          |
| p-value       |          |          |          | 0.0181   |          |
| Google 24 March|        |          |          | 30.39    |          |
| p-value       |          |          |          | 0.0251   |          |

### TABLE 4 Models of the number of deaths on 24 March 2020 using Google searches by state

| Deaths | 3 March | 10 March | 17 March | 20 March | 24 March |
|--------|---------|----------|----------|----------|----------|
| $R^2$  | 0.1315  | 0.1597   | 0.1027   | 0.0933   | 0.0785   |
| Adjusted $R^2$ | 0.1134  | 0.1422   | 0.0841   | 0.0774   | 0.0593   |
| Intercept | −18.87  | −26.26   | −43.51   | −51.03   | −45.9    |
| p-value | 0.1188  | 0.0497   | 0.0751   | 0.0804   | 0.118    |
| Google 3 March| 0.5879  |          |          |          |          |
| p-value | 0.0096  |          |          |          |          |
| Google 10 March|        | 0.6573   |          |          |          |
| p-value | 0.004   |          |          |          |          |
| Google 17 March|        |          | 0.7092   |          |          |
| p-value | 0.0232  |          |          |          |          |
| Google 20 March|        |          |          | 0.7774   |          |
| p-value |          |          |          | 0.031    |          |
| Google 24 March|        |          |          |          | 0.7155   |
| p-value |          |          |          |          | 0.0486   |

1The Google searches are a relative measure and the Wikipedia page lookups are actual numbers. The graph for Wikipedia is similar.
These results suggest that the searches, tweets, and page views occurred in three distinct waves. We can anticipate that, in the future, additional waves would also be embedded in the searches and page views. This view of waves is consistent with recent news developments (Coyne, 2020).

5 | CONCLUSION

Information gathered from the unintentional crowd may help provide guidance for the management of the current or a future pandemic. Our analysis found that models with lagged variables of the number of Google searches, Twitter tweets, and Wikipedia page views provide a leading indicator of the number of coronavirus cases and the number of deaths. Such a model could help provide government and hospital officials with important planning information.

It is unclear how long the coronavirus pandemic will last or if it will return after it passes. Since the coronavirus outbreak is only one of a set of related outbreaks that includes severe acute respiratory syndrome and Middle East respiratory syndrome, it is likely that we will witness additional virus outbreaks in the future (Haseltine, 2020). Further, as noted by Baumgartner and Rainey (2020), a recent US programme had identified over 1,200 viruses with the potential to cause a pandemic. Thus, it is important to continue to study how to incorporate Google searches, Twitter tweets, and Wikipedia page searches as part of models of disease diffusion in order to accurately anticipate and prepare for the potential number of cases and deaths.

5.1 | Challenges to Internet-Use-Generated Data

Some analysts and observers expect that Internet-based data, as used in this paper, would not necessarily provide insight into infectious diseases, for several reasons. Google Trends, Twitter, and Wikipedia do not consider challenges such as the digital divide. Further, in the current pandemic, it is already becoming clear that a high number of cases and deaths are recorded among vulnerable old populations or minority communities, who arguably tend to use digital tools much less. Further, researchers such as Arora, McKee and Stuckler (2019) argue that Google Trends may confound terms and is at a high level of granularity. We agree that these are important points in forecasting both number of coronavirus cases and deaths. However, we also argue that each of those concerns would work against our results, negatively influencing fit. As a result, perhaps the best results would be generated by adding these variables to existing models (e.g. Osthus, Daughton, and Priedhorsky, 2019).

5.2 | Potential Extensions

Although our work investigated the diffusion of the coronavirus across the 50 US states and the District of Columbia, a similar approach could be used to analyse either more aggregated country data or less aggregated city or county data. Similarly, we have focused largely on English resources, but the same approach could be applied to non-English resources. Further, other data sources are becoming available that could be incorporated. For example, there

[2]We thank one of the anonymous referees for posing these concerns.
has been recent discussion of a thermometer with an app that captures information about temperature, robots that spot people not wearing a mask, and helmets that find people with an elevated temperature within 5 meters of the helmet wearer (Maddow, 2020; Jakhar, 2020). Analysing such information is likely to provide additional data with the capabilities to predict both cases and deaths. Finally, as the pandemic continues, additional data will be added that can be used to test the results from early in the pandemic life cycle to those found later in the life cycle. We will be able to compare models built from early data, as seen in this paper, with models built on later data.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous referees for their comments on an earlier version of this paper. This research was supported by the J. Mack Robinson College of Business, Georgia State University.

REFERENCES

Althouse, B. M., Scarpino, S. V., Meyers, L. A., Ayers, J. W., Bargsten, M., Baumbach, J., Del Valle, S. (2015). Enhancing disease surveillance with novel data streams: challenges and opportunities. EPJ Data Science, 4(1), 1–8.

Aramaki, E., Maskawa, S., Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. In EMNLP ’11: Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1568–1576). Stroudsburg, PA: Association for Computational Linguistics.

Arora, V. S., McKee, M., Stuckler, D. (2019). Google Trends: opportunities and limitations in health and health policy research. Health Policy, 123(3), 338–341. https://doi.org/10.1016/j.healthpol.2019.01.001.

Banda, J., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Chowell, G. (2020). A Twitter dataset of 150+ million tweets related to COVID-19 for open research, April 3, https://zenodo.org/record/3738018#.XpJKuCSlAQ

Bardak, B., & Tan, M. (2015). 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE). Piscataway, NJ: IEEE. https://doi.org/10.1109/BIBE.2015.7367640

Baumgartner, E., & Rainey, J. (2020). U.S. ended program to find deadly viruses. Los Angeles Times, 3 April; A1, A6.

Butler, D. (2013). When Google got flu wrong: US outbreak foaxes a leading web-based method for tracking seasonal flu. Nature, 494(7436), 155–157. https://doi.org/10.1038/494155a

Cohen, J. (2020). Scientists are racing to model the next moves of a coronavirus that’s still hard to predict. Science. https://www.sciencemag.org/news/2020/02/scientists-are-racing-model-next-moves-coronavirus-thats-still-hard-predict

Coyne, M. (2020). The U.S. may be heading to a second (or even fourth) wave of the coronavirus epidemic. Here’s what that means. Forbes Magazine. https://www.forbes.com/sites/marleycoyne/2020/04/02/the-us-may-be-heading-to-a-second-or-even-fourth-wave-of-the-coronavirus-epidemic-heres-what-that-means/

Fleet J. (2020). A letter home from Shanghai. COVID-19 epidemic update, what do we do now edition. CGTN. https://news.cgtn.com/news/2020-03-13/A-letter-home-from-Shanghai-what-do-we-do-now-edition-OOXuGnML/index.html

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012–1014. https://doi.org/10.1038/nature07634

Google, 2015. The next chapter for flu trends. Google AI Blog. https://ai.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html

Haselton, W., (2020). Want to prevent another coronavirus epidemic? Scientific American. January 29, https://blogs.scientificamerican.com/observations/want-to-prevent-another-coronavirus-epidemic/

Jakhar, P. (2020). Coronavirus: China’s tech fights back. BBC News. https://www.bbc.com/news/technology-51717164

Kugler, L. (2016). What happens when big data blunders? Communications of the ACM, 59, 15–16. https://doi.org/10.1145/2911975

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. Science, 343(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Maddow, R., 2020. Transcript: The Rachel Maddow Show, http://www.msnbc.com/transcripts/rachel-maddow-show/2020-03-18

McIver, D. J., Brownstein, J. S. (2014). Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Computational Biology, 10(4), e1003581. https://doi.org/10.1371/journal.pcbi.1003581

Milinovich, G. J., Williams, G. M., Clements, A. C., & Hu, W. (2014). Internet-based surveillance systems for monitoring emerging infectious diseases. The Lancet. Infectious Diseases, 14(2), 160–168.

Nutl, S. V., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R. P., Chen, S. I., & Murugiah, K. (2014). The use of Google Trends in health care research: a systematic review. Plos One, 9(10), e109583.

O’Leary, D. E. (1998). Knowledge acquisition from multiple experts: An empirical study. Management Science, 44(8), 1049–1058.

O’Leary, D. E. (2015). Twitter mining for discovery, prediction and causal- ity: applications and methodologies. Intelligent Systems in Accounting, Finance and Management, 22(3), 227–247. https://doi.org/10.1002/isaf.1376

O’Leary, D. E. (2020). Evolving information systems and technology research issues for COVID-19 and other pandemics. Journal of Organizational Computing and Electronic Commerce, 30, 1–8.

Osthus, D., Daughton, A. R., & Friedhorsky, R. (2019). Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. PLoS Computational Biology, 15(2), e1006599. https://doi.org/10.1371/journal.pcbi.1006599

Pan, Y., & Jackson, R. T. (2008). Ethnic difference in the relationship between acute inflammation and serum ferritin in US adult males. Epidemiology and Infection, 136(03), 421–431. https://doi.org/10.1017/S095026880700831X

Paul, M. J., & Dredze, M. (2011). Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (pp. 265–272). Menlo Park, CA: AAAI Press.

Priedhorsky, R., Osthus, D., Daughton, A. R., Moran, K. R., Generous, N., Fairchild, G., Del Valle, S. Y. (2017). Measuring global disease with Wikipedia: success, failure, and a research agenda. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 1812–1834). New York, NY: ACM.

Renken, E., & Wood, D. (2020). Map: tracking the spread of the coronavirus in the U.S. NPR. https://www.npr.org/sections/health-shots/2020/03/16/816707182/map-tracking-the-spread-of-the-coronavirus-in-the-u-s

Ritterman, J., Osborne, M., & Klein, E. (2009) November. Using prediction markets and Twitter to predict a swine flu pandemic. In 1st international workshop on mining social media (Vol. 9, pp. 9–17). https://core.ac.uk/download/pdf/28976526.pdf

Sharpe, J. D., Hopkins, R. S., Cook, R. L., & Striley, C. W. (2016). Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis. JMIR Public Health and Surveillance, 2(2), e161.

Tausczik, Y., Faasse, K., Pennebaker, J. W., & Petrie, K. J. (2012). Public anxiety and information seeking following the H1N1 outbreak: blogs,
newspaper articles, and Wikipedia visits. *Health Communication*, 27(2), 179–185. https://doi.org/10.1080/10410236.2011.571759

World Health Organization. (2020). Coronavirus. https://www.who.int/health-topics/coronavirus#tab=tab_1

**How to cite this article:** O'Leary DE, Storey VC. A Google–Wikipedia–Twitter Model as a Leading Indicator of the Numbers of Coronavirus Deaths. *Intell Sys Acc Fin Mgmt*. 2020;27:151–158. https://doi.org/10.1002/isaf.1482