medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Internet Search Patterns Reveal Clinical Course of Disease Progression for COVID-19 and Predict Pandemic Spread in 32 Countries

Tina Lu[1,2], Ben Y. Reis* [1,3]

1. *Predictive Medicine Group, Computational Health Informatics Program, Boston Children's Hospital, Boston, MA*
2. *Harvard University, Cambridge, MA*
3. *Harvard Medical School, Boston, MA*

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Effective public health response to COVID-19 relies on accurate and timely surveillance of local pandemic spread, as well as rapid characterization of the clinical course of disease in affected individuals. *De novo* diagnostic testing methods developed for emergent pandemics are subject to significant development delays and capacity limitations. There is a critical need for complementary surveillance approaches that can function at population-scale to inform public health decisions in real-time. Internet search patterns provide a number of important advantages relative to laboratory testing. We conducted a detailed global study of Internet search patterns related to COVID-19 symptoms in multiple languages across 32 countries on six continents. We found that Internet search patterns reveal a robust temporal pattern of disease progression for COVID-19: Initial symptoms of fever, dry cough, sore throat and chills are followed by shortness of breath an average of 5.22 days [95% CI 3.30-7.14] after symptom onset, matching the precise clinical course reported in the medical literature. Furthermore, we found that increases in COVID-19-symptom-related searches predict increases in reported COVID-19 cases and deaths 18.53 days [95% CI 15.98-21.08] and 22.16 days [95% CI 20.33-23.99] in advance, respectively. This is the first study to show that Internet search patterns can be used to reveal the detailed clinical course of a disease. These data can be used to track and predict the local spread of COVID-19 before widespread laboratory testing becomes available in each country, helping to guide the current public health response.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Accurate real-time surveillance of local disease spread is essential for effective pandemic response, informing key public health measures such as social distancing and closures, as well as the allocation of scarce healthcare resources such as ventilators and hospital beds.[1 2] During the current COVID-19 pandemic, surveillance has primarily relied on direct testing of individuals using a variety of *de novo* laboratory testing methods developed specifically for the emergent pandemic.[3]

While laboratory testing serves as an important gauge of epidemic spread, it suffers from a number of important limitations: 1. During the early stages of a novel pandemic, laboratory tests specific to the pandemic do not yet exist and therefore must be developed *de novo*, leading to significant delays before test-based surveillance can begin. 2. Even once a test has been developed, scaling of manufacturing, distribution and test processing capacity takes a significant amount of time, resulting in limited testing capacity in many countries.[4] It is therefore difficult to achieve population-scale surveillance with laboratory testing in the crucial early stages of an emergent pandemic. 3. Even after tests are widely available, delays in test administration and processing make laboratory testing a lagging indicator relative to disease onset.[5] 4. Laboratory testing often requires individuals to leave home and congregate at testing centers, increasing exposure both for those being tested and for the health workers administering the tests. 5. Laboratory tests are typically invasive, involving blood draws or deep nasal swabs. 6. Laboratory testing can be expensive, especially when testing large numbers of people.[5]

Alternative surveillance approaches are needed to overcome these limitations and serve as a complement to laboratory testing, especially during the critical early stages of a pandemic. Aggregated de-identified Internet search patterns have been used to track a wide range of health phenomena, including influenza[6], MERS[7], measles[8], abortion[9] and immunization compliance,[10] and are a potential alternative source of information for surveilling pandemic spread. Previous uses of these data have yielded valuable lessons in their appropriate use, with an emphasis on avoiding non-specific search terms and avoiding complex models that tend towards overfitting.[11 6 12] When harnessed appropriately, Internet search patterns possess a number of powerful advantages relative to laboratory testing: 1. Surveillance data are available immediately when a new pandemic emerges. 2. Data are available at population-scale in countries with sufficient Internet access levels. [13] 3. Delays are minimal, with search data available the same day. 4. There is no need for individuals to travel to a testing location; people can stay at home, avoiding

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

increased exposure to themselves and to healthcare workers. 5. No physical intervention is required. 6. The data are available for free, independent of the scale of surveillance.

Another priority in the early stages of an emergent pandemic such as COVID-19 is to characterize the clinical course of symptoms in affected individuals. A number of recent clinical case studies have analyzed the day-by-day clinical course for COVID-19 patients: A study of 138 patients in Wuhan, China found that common symptoms include fever and dry cough, and the median time from initial symptom onset to dyspnea was 5.0 days.[14] A recent meta-review found that the average length from illness onset to dyspnea was 4.99 days, based on evidence from 179 COVID-19 patients.[15] A study of 41 patients in Wuhan, China found that the median duration from illness onset to dyspnoea was 8.0 days (IQR 5.0–13.0).[16] The CDC's *Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19)* states "Among patients who developed severe disease, the medium time to dyspnea ranged from 5 to 8 days".[17] These studies relied on a limited number of patients, and were published weeks and months after the initial spread of the pandemic. It would thus be beneficial to pandemic tracking, case diagnosis and treatment if these clinical patterns could be ascertained earlier and at population-scale.

We conducted a detailed global study across 32 countries on six continents to determine whether Internet search patterns can provide reliable real-time indicators of local COVID-19 spread, and whether these data can reveal the clinical progression of COVID-19.

Methods

We selected 32 countries from diverse regions of the world (Table 1), in which robust search data were available for the search terms of interest. We obtained data on reported COVID-19 cases and deaths for each of these countries from a publicly available dataset maintained by the Center for Systems Science and Engineering at Johns Hopkins University. [18]

We collected daily relative search volume (RSV) data on a per-country basis for the period of January 1, 2020 through April 20, 2020, from Google Trends[19] using the pytrends API [20]. Since Google has limited availability in China, we accessed search trend data from the search engine Weibo.[21] We accessed data for the following common symptoms of COVID-19: "fever", "cough",

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

"dry cough", "chills", "sore throat", "runny nose" and "shortness of breath", as well as the general terms "coronavirus", "coronavirus symptoms" and "coronavirus test". We also looked at other less common symptoms such as loss of smell and loss of taste, but the search data on those terms were too sparse for many countries. Since daily search data are inherently noisy, all search data were smoothed with a 5-day moving average.

For translations of the English search terms into local languages, we recruited native speakers to translate each term into the following languages: Arabic, Mandarin Chinese, Dutch, French, German, Italian, Persian, Polish, Portuguese and Spanish. These languages covered almost all of the countries included in the study. For languages in which we were unable to recruit native speakers for translations, we used Google Translate[22] and confirmed that robust data were available for these translated search terms on Google Trends.[19]

We conducted temporal correlation studies to characterize the relationships between search data and reported COVID-19 cases and deaths. For each country and search term, we calculated the Pearson correlation coefficient between the time series of search volumes for that search term and the time series of COVID-19 cases. We then shifted the search term data by a variable lag, and identified the lag that yielded the highest correlation. We computed the mean and standard deviation of these optimal lags for each search term across all countries. We then repeated these analyses substituting reported COVID-19 deaths for reported COVID-19 cases.

We then investigated whether Internet search data can be used to characterize the clinical course of COVID-19 symptoms over time. We used "coronavirus symptoms" as the index search term, since it peaked first among all other search terms. For each country, the date of peak search volume for the index search term was defined as the index date, and the dates for all other search data for that country were defined in relation to this date (Day 0, Day 1, Day 2, etc.). With the data from all countries thus aligned, cross-country ensemble average curves were calculated for each search term. (e.g. Day 1 values for "fever" searches in each of the 32 countries were averaged together to calculate the Day 1 value of the ensemble average "fever" curve. The same was done for Day 2, etc.) The ensemble average curves for all search terms were then overlaid on one plot, providing a search-data-based view of the clinical course of illness. As above, all results were calculated using search data smoothed with a 5-day moving average. For

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

visualization purposes a 10-day moving average provided slightly clearer plots; plots based on 5-day moving average are included in the Supplementary Materials.

Results

We begin by presenting examples from individual countries. Figure 1 shows search volumes for the terms "fever" and "dry cough", alongside reported COVID-19 cases and deaths for China, Iran, Italy, United States and India. Even though outbreaks occurred at different times in each country, the temporal relationships between the search terms and reported COVID-19 cases and deaths are similar across countries.

Figure 2a shows the lags between search volumes for the term "fever" and COVID-19-related deaths across 32 countries, along with a histogram showing the distribution of these lags. Figure 2b shows the same information for the term "dry cough". Table 2 shows summary statistics for the lags between each of the 10 search terms and COVID-19 cases and deaths, including mean, 95% confidence intervals and standard deviation across 32 countries. Searches for "fever" predict increases in reported COVID-19 cases and deaths 18.53 days [95% CI 15.98-21.08] and 22.16 days [95% CI 20.33-23.99] in advance, respectively. For "cough" these times were 18.34 days [95% CI 16.05-20.64] and 21.88 days [95% CI 20.19-23.56], respectively.

The average lags between searches and reported cases are shorter than those between searches and reported deaths since cases are usually diagnosed and reported before deaths. The cross-country variability of the average lags between searches and reported cases is greater (higher standard deviation and larger confidence intervals) than that for reported deaths, likely since case reporting is more sensitive to local testing capacity and rates which vary greatly between countries.

We found that the general term "coronavirus" has greater variability in its lags to reported cases and deaths (largest standard deviation and confidence intervals), compared to other symptom-specific terms, as would be expected for a non-specific search term.

We examined the progression of symptom-related search terms over time in order to characterize the clinical course of COVID-19. Figure 3 shows examples from individual

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

countries, with search volumes for a range of symptom-related search terms plotted over time for both France and Mexico. In each country, searches for "shortness of breath" appeared a few days after searches for other symptoms. (Figure S-1 in the Supplementary materials shows the same results smoothed with a 5-day moving average.)

We calculated ensemble average curves for each search term across 32 countries. Figure 4a shows the ensemble average search volumes for "fever", "cough", "dry cough" and "shortness of breath", indexed by searches for "coronavirus symptoms", alongside reported COVID-19 cases and deaths. Figure 4b shows this same analysis for additional search terms "sore throat", "runny nose", "chills", and "coronavirus test", also indexed by searches for "coronavirus symptoms". (Figure S-2 in the Supplementary materials shows the same results smoothed with a 5-day moving average.)

The clinical progression that emerges from these data presents the following picture: As the pandemic begins to take hold in a country, people search for "coronavirus symptoms" and "coronavirus test", followed by initial symptoms "fever", "cough", "runny nose", "sore throat" and "chills", followed by searches for "shortness of breath" about 5 days after the search for initial symptoms.

The medical literature reports the clinical progression of COVID-19 in terms of the number of days between initial symptom onset and shortness of breath (dyspnea). Therefore, we examined a range of possible search-term-based definitions for initial symptom onset, based on various combinations of the earliest-peaking search terms "fever", "cough", "coronavirus symptoms" and "coronavirus test". Table 3 shows the lags between these different definitions for initial symptom onset and searches for "shortness of breath". The average lag between the searches for "fever" and "shortness of breath" was 5.22 days [95% CI 3.30-7.14]. For "cough" it was 5.16 days [95% CI 3.13-7.18]. These lags, and lags deriving from other symptom onset definitions, are all around 5 days, precisely matching the clinical course of the disease reported in the literature.[14 15 16 17]

Discussion

This is the first study to show that Internet search patterns can be used to reveal the detailed clinical progression of a disease. During emergent pandemics, this level of detail can provide

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

public health officials not only with predictions of local pandemic spread, but also with a detailed understanding of the stage of illness and the manifestations of the disease in the local environment. Recent studies have indicated that the spread and severity of the disease can be affected by local conditions,[23 24 25] and search volume data can be a valuable complementary tool in studying potential local variations in disease presentation.

We found that increases in symptom-related searches such as "fever" predict increases in reported COVID-19 cases and deaths 18.53 days [95% CI 15.98-21.08] and 22.16 days [95% CI 20.33-23.99] in advance, respectively. This robust relationship across countries and languages, allows public health officials to monitor expected pandemic spread through Internet search activity even before local testing is available. At the time of manuscript preparation, search volumes for COVID-19 related symptoms are still increasing in a number of countries, and these data can be useful in gauging real-time spread.

The use of search data is subject to a number of important limitations.[11 6 12] Internet infrastructure and digital access levels differ across countries and communities. Some countries currently lack sufficient search volume to support robust search-based tracking, though in the long-term, digital access rates are increasing worldwide. Search data may be subject to socio-economic, geographic, or other biases inherent in the local digital divide.[13,26,27] Even though selecting specific search terms increases the signal to noise ratio, changes in search volumes for symptom-related terms such as "fever" could result not only from increases in COVID-19 cases, but also general curiosity about the pandemic, other diseases (e.g. influenza, Lassa fever[28]), news coverage, or other factors. Due to privacy considerations, search data are provided as aggregated relative search volumes rather than raw counts, so while it is possible to predict an increase in cases, it may be difficult to infer the magnitude of such an increase. While search data have important limitations, laboratory testing is subject to a wide range of limitations listed above, in addition to reliability issues which cause certain types of laboratory tests to need to be repeated in order to increase reliability.[29 30]

Recent studies have examined Internet search data related to the current COVID-19 pandemic. Some studies looked only at the search term "coronavirus",[31, 32, 33, 34] while others looked at additional search terms such as "handwashing", "face masks",[35 36] "quarantine", "hand disinfection",[37] "SARS", "MERS", [38] "antiseptic", and "sanitizer".[39] None of these studies

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

examined specific symptom-related search terms. In our analysis, we found that general non-symptom-specific search terms such as "coronavirus" have a greater variability in their relation to reported cases and deaths. This is likely due to the fact that individuals seeking general information on the pandemic may search for "coronavirus" even if they are not experiencing specific symptoms themselves at that time.

A few studies have examined Internet search data using terms related to COVID-19 symptoms, but these covered only one country or a small number of countries: One study looked at "coronavirus" and "pneumonia" only in China and found that Internet searches were correlated with daily incidences of COVID-19.[40] Another looked at "COVID", "COVID pneumonia", and "COVID heart" only in the US, and found that these terms were correlated with COVID-19 daily incidence and deaths.[41] Yet another study looked at "loss of smell" in 8 countries, and found these to be correlated with COVID-19 cases.[42] Another study developed a model incorporating a large selection of search terms, including some symptom-related terms, in 8 countries; however this latter study focused on optimizing the performance of a predictive model, rather than studying the detailed temporal relationships between patterns of symptoms.[43]

The present study is the first to conduct a detailed investigation of multiple COVID-19 symptom-related search terms across a large number of countries. It is also the first to conduct a detailed analysis of the temporal relationships between different symptom-related searches in order to characterize the clinical course of illness for COVID-19. Future work includes training a robust predictive model with various machine learning techniques to provide more granular predictions for increases in COVID-19 cases and deaths. Such models can also take into account additional sources of data such as news reports, testing capacity, public health mitigation measures, climatological and air quality variables, among others.

The ability of search data to not only predict future increases in cases, but also reveal the clinical course of symptoms in emergent pandemics is significant. Given the numerous limitations of laboratory testing, search data are a valuable complementary source for population-scale tracking of pandemics in real time. These data can be used today to guide the public health response to the COVID-19 pandemic in countries worldwide.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Data Availability

All data used in this study are publicly available through the sources referenced in the Methods section.

Author Contributions

B.Y.R. and T.L conceived of the study, supervised its conduct, and oversaw data collection. B.Y.R. and T.L. designed the study, conducted statistical analysis, drafted the manuscript and formulated the implications of the results. All authors contributed substantially to the revision of the manuscript and approved the final manuscript as submitted.

Competing Interest Statement

The authors declare no competing interests.

References

1. Fineberg, H. V. Pandemic preparedness and response--lessons from the H1N1 influenza of 2009. *N. Engl. J. Med.* **370**, 1335–1342 (2014).
2. Balter, S. *et al.* Pandemic (H1N1) 2009 surveillance for severe illness and response, New York, New York, USA, April-July 2009. *Emerg. Infect. Dis.* **16**, 1259–1264 (2010).
3. Peto, J. Covid-19 mass testing facilities could end the epidemic rapidly. *BMJ* **368**, (2020).
4. Godlee, F. The burning building. *BMJ* **368**, (2020).
5. Sheridan, C. Fast, portable tests come online to curb coronavirus pandemic. *Nat. Biotechnol.* (2020) doi:10.1038/d41587-020-00010-2.
6. Kandula, S. & Shaman, J. Reappraising the utility of Google Flu Trends. *PLoS Comput. Biol.* **15**, e1007258 (2019).
7. Shin, S.-Y. *et al.* High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. *Sci. Rep.* **6**, 32920 (2016).
8. Santangelo, O. E. *et al.* Digital epidemiology: assessment of measles infection through Google Trends mechanism in Italy. *Ann. Ig.* **31**, 385–391 (2019).
9. Reis, B. Y. & Brownstein, J. S. Measuring the impact of health policies using Internet search patterns: the case of abortion. *BMC Public Health* **10**, 514 (2010).
10. Barak-Corren, Y. & Reis, B. Y. Internet activity as a proxy for vaccination compliance. *Vaccine* vol. 33 2395–2398 (2015).
11. Arora, V. S., McKee, M. & Stuckler, D. Google Trends: Opportunities and limitations in health and health policy research. *Health Policy* **123**, 338–341 (2019).
12. Lazer, D., Kennedy, R., King, G. & Vespignani, A. Big data. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
13. Alicino, C. *et al.* Assessing Ebola-related web search behaviour: insights and implications

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

from an analytical study of Google Trends-based query volumes. *Infect Dis Poverty* **4**, 54 (2015).

14. Wang, D. *et al.* Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA* vol. 323 1061 (2020).
15. Tan, S. C. Clinical and epidemiological characteristics of Coronavirus Disease 2019 (COVID-19) patients. doi:10.1101/2020.04.02.20050989.
16. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
17. Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html.
18. CSSEGISandData. CSSEGISandData/COVID-19. *GitHub* https://github.com/CSSEGISandData/COVID-19.
19. Website. https://trends.google.com/trends/explore.
20. GeneralMills. GeneralMills/pytrends. *GitHub* https://github.com/GeneralMills/pytrends.
21. Weibo Search Trends. https://data.weibo.com/index.
22. Google Translate. https://translate.google.com/.
23. Wang, J., Tang, K., Feng, K. & Lv, W. High Temperature and High Humidity Reduce the Transmission of COVID-19. *SSRN Electronic Journal* doi:10.2139/ssrn.3551767.
24. Mecenas, P., Bastos, R., Vallinoto, A. & Normando, D. Effects of temperature and humidity on the spread of COVID-19: A systematic review. doi:10.1101/2020.04.14.20064923.
25. Wu, X., Nethery, R. C., Sabath, B. M., Braun, D. & Dominici, F. Exposure to air pollution and COVID-19 mortality in the United States. doi:10.1101/2020.04.05.20054502.
26. Din, H. N., McDaniels-Davidson, C., Nodora, J. & Madanat, H. Profiles of a Health Information-Seeking Population and the Current Digital Divide: Cross-Sectional Analysis of the 2015-2016 California Health Interview Survey. *J. Med. Internet Res.* **21**, e11931 (2019).
27. McCloud, R. F., Okechukwu, C. A., Sorensen, G. & Viswanath, K. Beyond access: barriers to internet health information seeking among the urban poor. *J. Am. Med. Inform. Assoc.* **23**, 1053–1059 (2016).
28. Wogu, J. O. *et al.* Mass media reportage of Lassa fever in Nigeria: a viewpoint. *J. Int. Med. Res.* **48**, 300060518821552 (2020).
29. Li, Z. *et al.* Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis. *J. Med. Virol.* (2020) doi:10.1002/jmv.25727.
30. Beeching, N. J., Fletcher, T. E. & Beadsworth, M. B. J. Covid-19: testing times. *BMJ* **369**, (2020).
31. Strzelecki, A. The second worldwide wave of interest in coronavirus since the COVID-19 outbreaks in South Korea, Italy and Iran: A Google Trends study. *Brain Behav. Immun.* (2020) doi:10.1016/j.bbi.2020.04.042.
32. Mavragani, A. Tracking COVID-19 in Europe: Infodemiology Approach. *JMIR Public Health Surveill* **6**, e18941 (2020).
33. Effenberger, M. *et al.* Association of the COVID-19 pandemic with Internet Search Volumes: A Google TrendsTM Analysis. *International Journal of Infectious Diseases* (2020)

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

doi:10.1016/j.ijid.2020.04.033.

34. Benjamin Lucas Brendan Elliot. Online Information Search During COVID-19. *arXiv:2004.07183v2* https://arxiv.org/pdf/2004.07183.pdf (2020).
35. Husnayain, A., Fuad, A. & Su, E. C.-Y. Applications of google search trends for risk communication in infectious disease management: A case study of COVID-19 outbreak in Taiwan. *Int. J. Infect. Dis.* (2020) doi:10.1016/j.ijid.2020.03.021.
36. Lin, Y.-H., Liu, C.-H. & Chiu, Y.-C. Google searches for the keywords of 'wash hands' predict the speed of national spread of COVID-19 outbreak among 21 countries. *Brain, Behavior, and Immunity* (2020) doi:10.1016/j.bbi.2020.04.020.
37. Jarynowski, A., Wojta-Kempa, M. & Belik, V. Perception of emergent epidemic of COVID-2019 / SARS CoV-2 on the Polish Internet. doi:10.1101/2020.03.29.20046789.
38. Strzelecki, A. & Rizun, M. Infodemiological Study Using Google Trends on Coronavirus Epidemic in Wuhan, China. *International Journal of Online and Biomedical Engineering (iJOE)* vol. 16 139 (2020).
39. Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M. & R Niakan Kalhori, S. Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study. *JMIR Public Health Surveill* **6**, e18828 (2020).
40. Li, C. *et al.* Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Euro Surveill.* **25**, (2020).
41. Yuan, X. *et al.* Trends and prediction in daily incidence and deaths of COVID-19 in the United States: a search-interest based model. doi:10.1101/2020.04.15.20064485.
42. Walker, A., Hopkins, C. & Surda, P. The use of google trends to investigate the loss of smell related searches during COVID-19 outbreak. *Int. Forum Allergy Rhinol.* (2020) doi:10.1002/alr.22580.
43. Vasileios Lampos, Simon Moura, Elad Yom-Tov, Michael Edelstein, Maimuna Majumder, Yohhei Hamada, Molebogeng X. Rangaka, Rachel A. McKendry, Ingemar J. Cox. Tracking COVID-19 using online search. *arXiv:2003.08086v7* https://arxiv.org/pdf/2003.08086.pdf (2020).

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Tables and Figures:

Region	Countries
Africa	Nigeria, South Africa, Zimbabwe
Asia	India, Iran, Malaysia, Saudi Arabia, Turkey
Oceania	Australia, Indonesia
North America	Canada, United States
Central America	Guatemala, Mexico
South America	Argentina, Brazil, Chile, Colombia, Uruguay, Venezuela
Europe	Belgium, France, Germany, Hungary, Ireland, Italy, Netherlands, Poland, Spain, Sweden, Switzerland, United Kingdom

Table 1. Countries included in the study, categorized by region.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Search Term	**Days to COVID-19 Cases** *Mean [95% CI] (Std. dev)*	**Days to COVID-19 Deaths** *Mean [95% CI] (Std. dev)*
"Coronavirus Symptoms"	21.97 [19.59-24.35] (6.60)	24.50 [22.79-26.21] (4.74)
"Coronavirus Test"	19.44 [17.12-21.76] (6.84)	22.66 [20.93-24.38] (4.78)
"Fever"	18.53 [15.98-21.08] (7.08)	22.16 [20.33-23.99] (5.07)
"Cough"	18.34 [16.05-20.64] (6.37)	21.88 [20.19-23.56] (4.68)
"Coronavirus"	17.53 [14.33-20.73] (8.87)	21.12 [18.63-23.62] (6.93)
"Runny Nose"	17.50 [14.71-20.29] (7.75)	21.06 [18.57-23.55] (6.90)
"Dry Cough"	17.44 [14.69-20.19] (7.63)	20.81 [19.09-22.52] (4.78)
"Sore Throat"	16.75 [14.22-19.28] (7.03)	20.84 [18.71-22.98] (5.93)
"Chills"	16.75 [14.12-19.38] (7.29)	19.38 [17.49-21.26] (5.24)
"Shortness of Breath"	13.81 [11.23-16.39] (7.15)	18.00 [15.77-20.23] (6.19)

Table 2. Lags between search volumes for specific search terms and reported COVID-19 cases and deaths across 32 countries. Mean lags are reported in days, along with 95% confidence intervals and standard deviations. Search terms that peaked earliest are listed first.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Symptom Onset *Search-Term-Based Definition*	**Days to "Shortness of Breath"** *Mean [95% CI] (Std. dev)*
"Fever"	5.22 [3.30-7.14] (5.32)
"Cough"	5.16 [3.13-7.18] (5.62)
Average of "Fever", "Cough"	5.19 [3.28-7.09] (5.29)
Average of "Fever", "Cough", "Coronavirus Symptoms"	5.81 [3.92-7.70] (5.24)
Average of "Fever", "Cough", "Coronavirus Symptoms", "Coronavirus Test"	5.71 [3.85-7.57] (5.16)

Table 3. Average lag in days from search-based symptom onset to searches for "Shortness of Breath" across 32 countries. Different search-term-based definitions for symptom onset were examined by looking at different combinations of early-peaking search terms.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

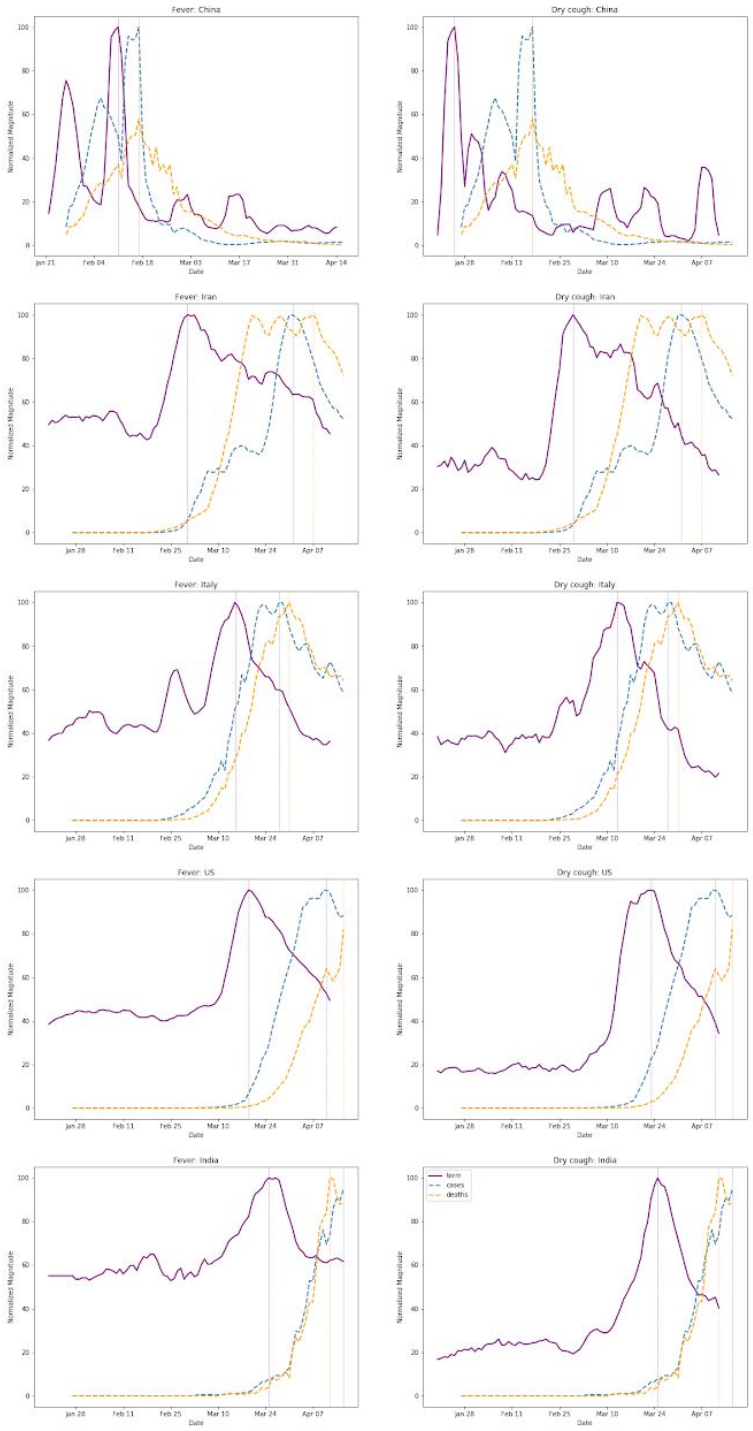
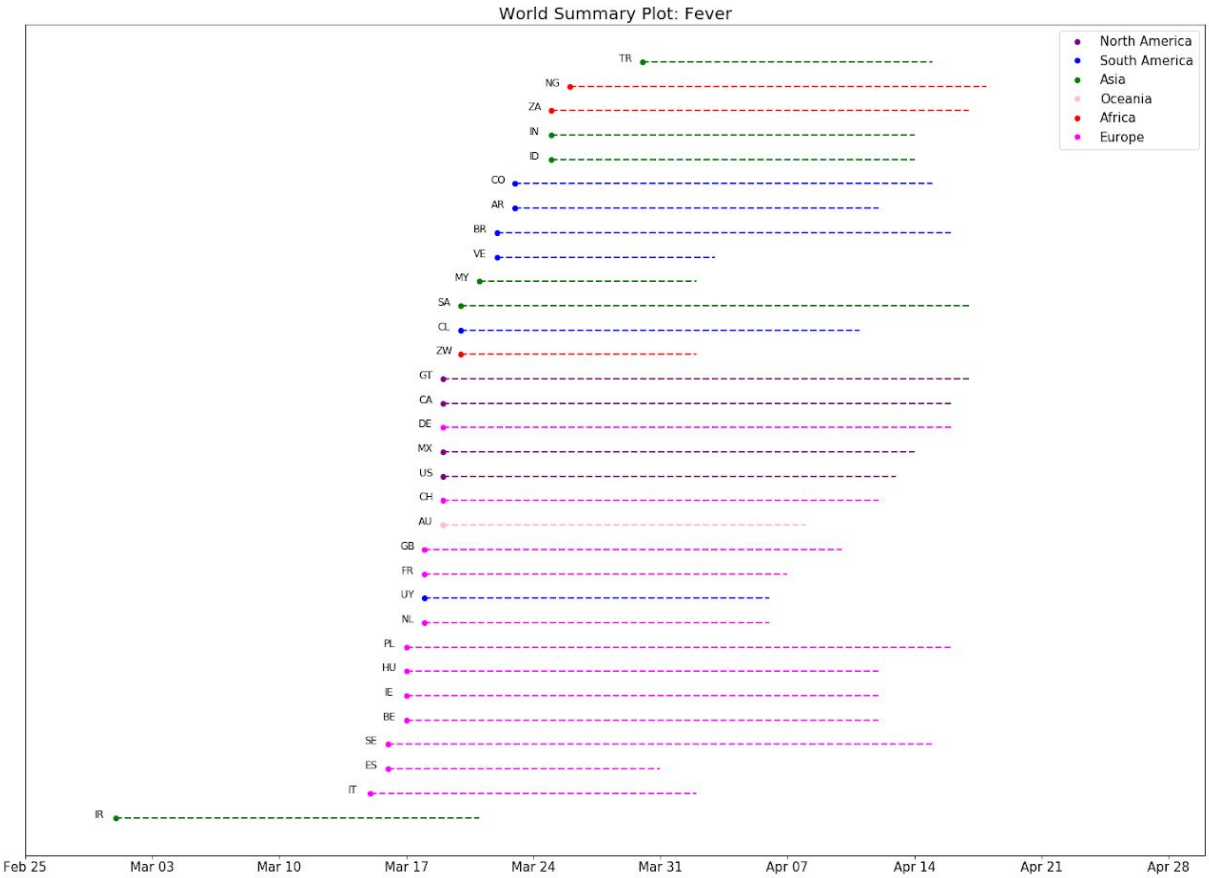
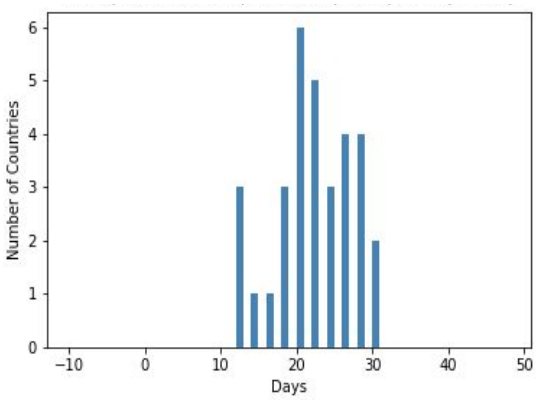


Figure 1. Search volumes (purple) for the terms "fever" (left) and "dry cough" (right), alongside reported COVID-19 cases (cyan) and deaths (orange) for China, Iran, Italy, US and India. Even though outbreaks occur at different times in different countries, the relationships between the search terms and reported COVID-19 cases and deaths are similar across countries. To highlight the temporal relationships between the curves, the magnitude of each curve was independently normalized to fit the vertical dimensions of the plot.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(a)





medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(b)

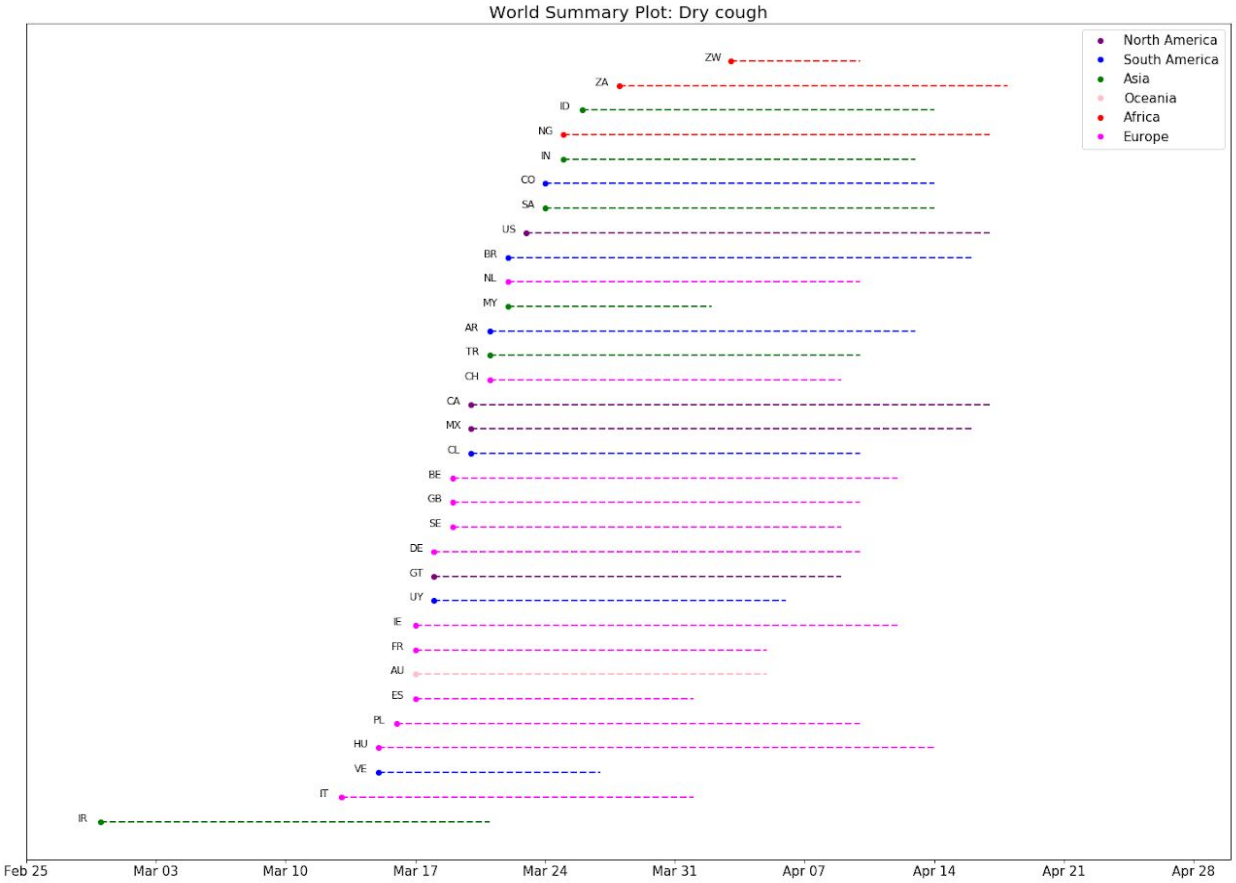


Figure 2. (a) Lags between searches for "fever" and reported COVID-19 deaths across 32 countries, with a histogram showing the distribution of these lags. Each country is labeled with its ISO Alpha-2 country code. **(b)** The same plots shown for searches for "dry cough".

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

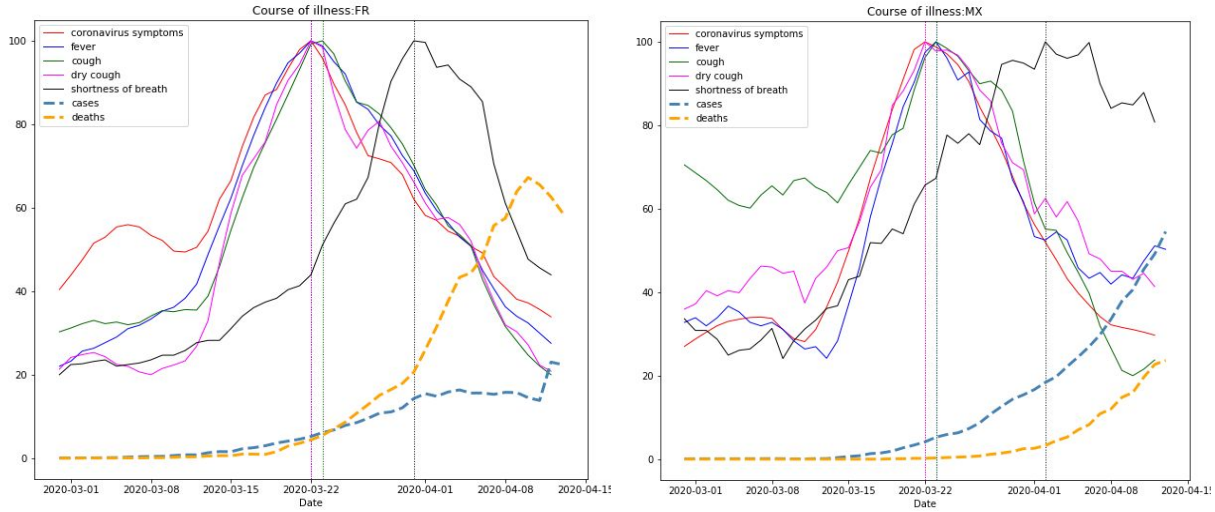


Figure 3. Clinical Course of COVID-19 symptoms in France (left) and Mexico (right). Search volumes for the terms “coronavirus symptoms” , “fever” , “cough”, “dry cough” and “shortness of breath” (black) are shown alongside COVID-19 cases (dashed cyan line) and deaths (dashed orange line). Initial symptoms appear clustered together in time, with searches for shortness of breath appearing a few days later. To highlight the temporal relationships between curves, the magnitude of each curve was independently normalized to fit the vertical dimensions of the plot.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

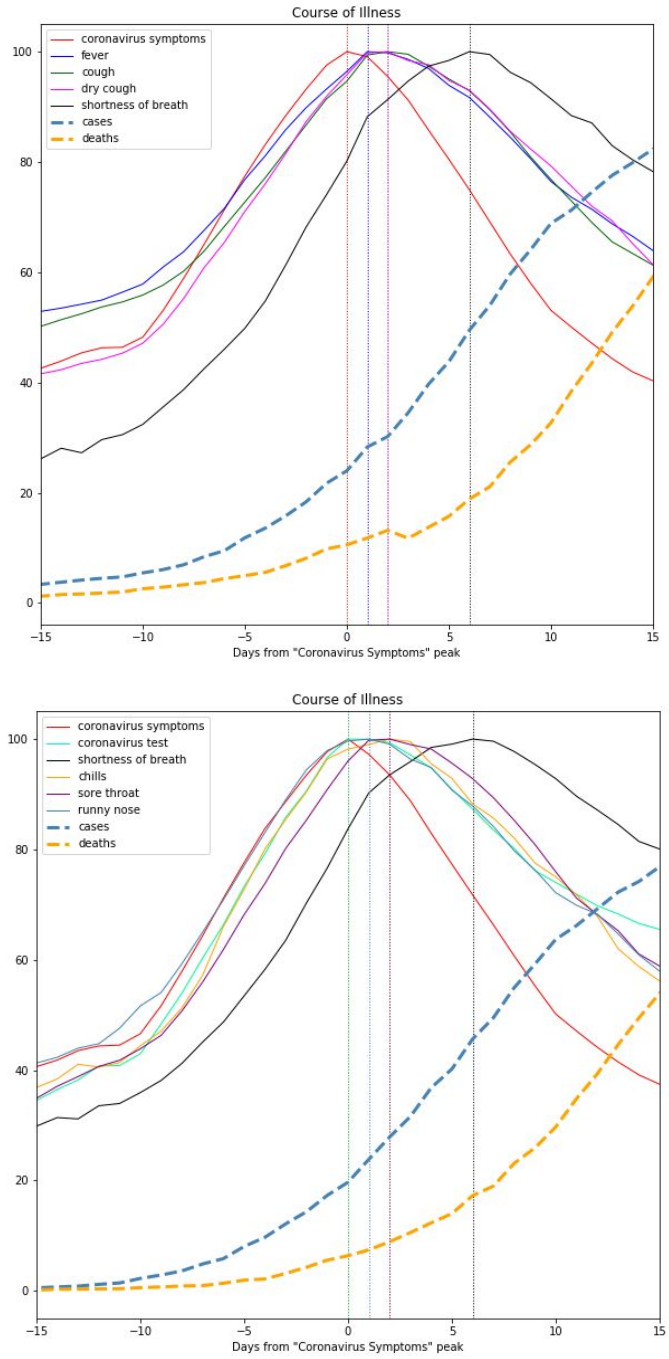


Figure 4. Ensemble average clinical course of COVID-19 symptoms across 32 countries, as seen through Internet search volumes of search terms related to COVID-19 symptoms. **(a)** Search volumes for the terms “fever” , “cough”, “dry cough”, “shortness of breath” (black), indexed by searches for “coronavirus symptoms”, shown alongside COVID-19 cases (dashed line cyan) and deaths (dashed line orange). **(b)** Search volumes for the terms “sore throat”, “runny nose”, “chills”, and “coronavirus test”, algonside “shortness of breath” (black), indexed by searches for “coronavirus symptoms”, shown alongside COVID-19 cases (dashed line cyan) and deaths (dashed line orange).

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Materials:

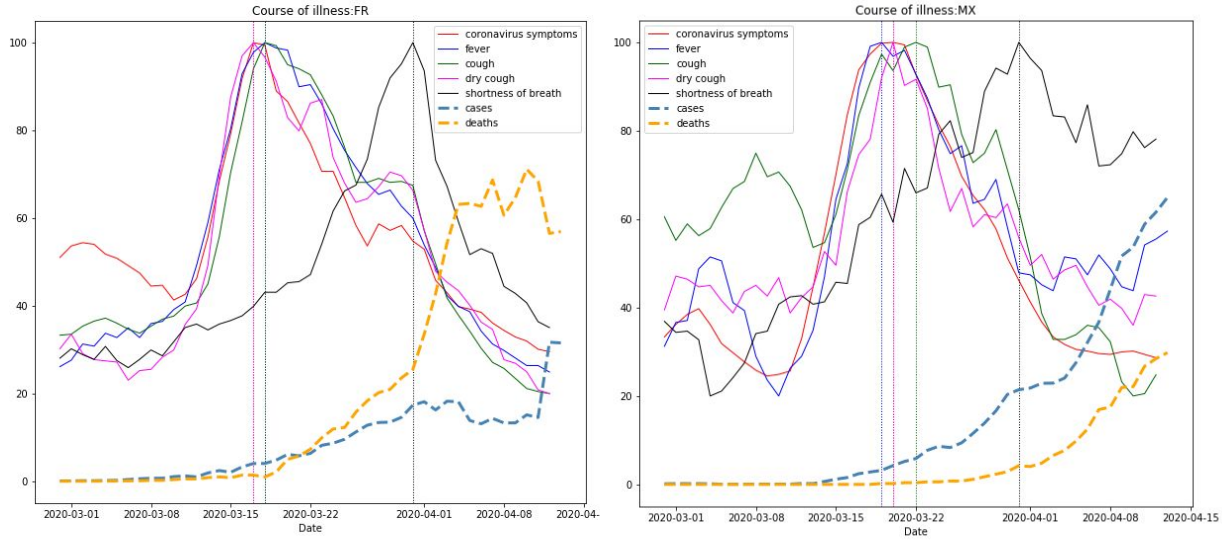


Figure S-1. Results from Figure 3, smoothed with a 5-day moving average rather than a 10-day moving average.

medRxiv preprint doi: https://doi.org/10.1101/2020.05.01.20087858.this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

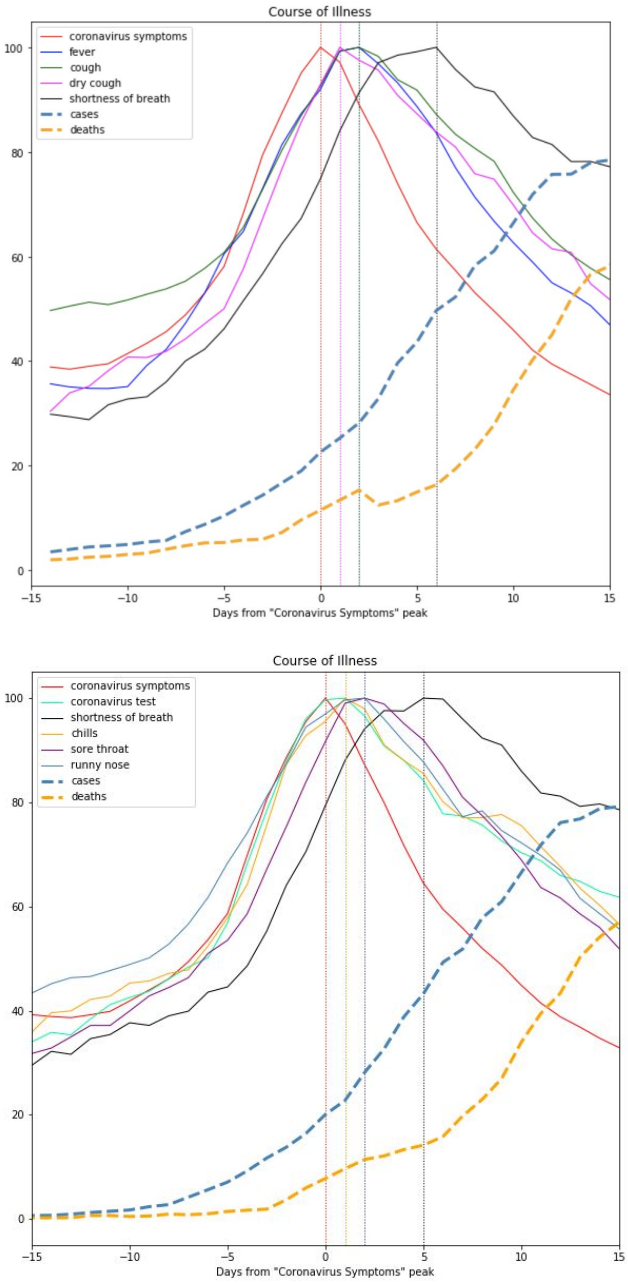


Figure S-2. Results from Figure 4, smoothed with a 5-day moving average rather than a 10-day moving average.