Analysis of hospital traffic and search engine data in Wuhan China indicates early disease activity in the Fall of 2019

Analysis of hospital traffic and search engine data in Wuhan China indicates early disease activity in the Fall of 2019

Elaine Okanyene Nsoesie[1]*, Benjamin Rader[2,3]*, Yiyao L. Barnoon[2], Lauren Goodwin[2], John S. Brownstein[2,4]

1. Department of Global Health, Boston University School of Public Health, Boston, USA
2. Computational Epidemiology Lab, Boston Children's Hospital, Boston, USA
3. Department of Epidemiology, Boston University School of Public Health, Boston, USA
4. Departments of Pediatrics and Biomedical Informatics, Harvard Medical School, Boston, USA

*authors contributed equally
Correspondence should be addressed to: john.brownstein@childrens.harvard.edu

Abstract:
The global COVID-19 pandemic was originally linked to a zoonotic spillover event in Wuhan's Huanan Seafood Market in November or December of 2019. However, recent evidence suggests that the virus may have already been circulating at the time of the outbreak. Here we use previously validated data streams - satellite imagery of hospital parking lots and Baidu search queries of disease related terms - to investigate this possibility. We observe an upward trend in hospital traffic and search volume beginning in late Summer and early Fall 2019. While queries of the respiratory symptom "cough" show seasonal fluctuations coinciding with yearly influenza seasons, "diarrhea" is a more COVID-19 specific symptom and only shows an association with the current epidemic. The increase of both signals precede the documented start of the COVID-19 pandemic in December, highlighting the value of novel digital sources for surveillance of emerging pathogens

Introduction

Early investigations into SARS-CoV-2 emergence and the resulting COVID-19 disease outbreak proposed the proximate cause was a zoonotic spillover event in late November or early December 2019 in Wuhan, China[1–3]. This was supported by preliminary epidemiological studies, including the initial clinical series which linked two-thirds of the identified cases to the Huanan Seafood Market in Wuhan[4,5]. Critically, the study found no direct connection to the market for 14 individuals, including the first known case of COVID-19, leaving open the possibility of alternate points of origin and infection[4]. Additionally, virologic samples of wildlife in the Huanan market could not be linked to SARS-CoV-2, suggesting transmission at the market was downstream from the spillover event[6]. Here we consider that SARS-CoV-2 may have already been circulating in the community prior to the identification of the Huanan Market cluster. This hypothesis is supported by emerging epidemiologic and phylogenetic evidence indicating that the virus emerged in southern China[7] and may have already spread internationally[8], and adapted for efficient human transmission[9] by the time it was detected in late December.

Digital epidemiology and non-traditional data streams, such as satellite imagery and internet search trends, have previously been harnessed for respiratory disease surveillance[10,11]. These sources have been shown to be early indicators of epidemics and sensitive to trends that may otherwise go undetected by traditional public health surveillance mechanisms[12,13]. In this study, we use two of these previously validated data streams to look for indicators of potential COVID-19 disease prior to December 2019. First, using vehicle counts extracted from high-resolution satellite imagery of hospital parking lots in Wuhan, we aim to estimate trends in hospital occupancy and its association with influenza-like illness (ILI) trends. This method has been demonstrated as an effective proxy for detecting hospital traffic related to respiratory illness in Latin America[11]. Second, we use Baidu search trends to try to determine the etiology of potential changes in ILI. We have previously shown that the volume of Baidu search queries can be used to estimate influenza trends in China[14]. Together, we assess whether these digital sources can augment traditional epidemiologic, genetic and virologic tools for understanding the timing of SARS-CoV-2 emergence.

Methods

Hospital traffic data

We obtained archived high-resolution satellite imagery (average resolution of about 70 cm) data for Wuhan, China from Remote Sensing Metrics (RS Metrics). We developed a comprehensive list of hospitals in Wuhan (using Google Maps, Wikipedia and PubMed). After the exclusion of sub-specialty hospitals (e.g. Wuhan Asia Heart) and hospitals with no satellite imagery available (Jinyintan), we identified 6 hospitals for imagery analysis: Hubei Women and Children's, Wuhan Tianyou, Wuhan Central, Wuhan Tongji Medical University, Wuhan Union, and Zhongnan Hospital of Wuhan University. We also identified high traffic areas including the Huanan Seafood Market and two railway stations (Wuchang and Hankou) for validation. For each location, RS Metrics identified cars in parking lots by first delineating hospital (or other site) premises, parking lot borders and street parking by automated feature extraction and then manual counting and quality control (as previously described)[11]. Images with tree cover, building

shadow, construction and other factors that present difficulties in defining the contours were excluded since this could lead to over- or under-counting of the number of vehicles. The process of data analysis was independent of the image selection process. The dataset used in analysis consisted of the date and time of each image, the hospital's name and geographic location (including the address, latitude and longitude), and the numbers of vehicles in the parking lot. A metric of relative daily car volume was computed: $relative\ count_{hj} = \frac{raw\ count_{hj}}{mean(raw\ count_{hs})}$, where for each hospital, *h,* and daily satellite image, *j*, counts were compared against baseline means for that hospital during each segment of the week, *s* (weekdays, Saturday, and Sunday). A loess (locally estimated scatterplot smoothing) regression line with a 40% smoothing span was fit to the data.

Search query data

Baidu's database (http://index.baidu.com/) contains logs of web and mobile search query volume in China. User confidentiality is maintained, since only the relative term frequency data is available. We obtained daily data for symptom-related searches likely associated with COVID-19 illness in Wuhan from April 2017 to May 2020. We extracted the relative search volumes of the terms "cough" and "diarrhea" using WebPlotDigitizer, v4.2[15].

Clinical Data

We obtained data on influenza-like illness from two sentinel hospitals in Wuhan: the Children's Hospital of Wuhan and Wuhan No. 1 Hospital, from Kong et al. (2020)[16]. The authors state that ILI trends noted in these two hospitals represent the overall trend in the local population. The two hospitals are the largest pediatric hospital in Hubei and a major general hospital, respectively. Counts of confirmed COVID-19 cases in Wuhan were aggregated from an open access repository of global line-list disease data[17].

Results

We collected 111 satellite images of Wuhan (multiple sites per image) from January 9, 2018 to April 30, 2020 resulting in 140 successful daily extractions of parking lot volume from hospitals (**Figure 1,** example on top panel) and 117 from the three high-volume control sites (**Figure 1,** example on bottom panel). Between 2018 and 2020, there was a general upward trend of increased hospital occupancy as measured by the parking lot volume proxy (**Figure 2, a**). The loess smoothed line shows a steep increase in volume starting in August 2019 (**Figure 2,** first annotation) and culminating with a peak in December 2019 (**Figure 2,** second annotation). Individual hospitals have days of high relative volume in both Fall and Winter 2019. However, between September and October 2019, 5 of the 6 hospitals show their highest relative daily volume of the analyzed series, coinciding with elevated levels of Baidu search queries for the terms "diarrhea" and "cough" (**Figure 2, b**). While searches for "diarrhea" only show elevated traffic starting in late 2019, "cough" shows yearly peaks that approximately coincide with influenza season (**Figure 2, c, turquoise**). Both search query terms show a large increase approximately 3 weeks preceding the large spike of confirmed COVID-19 cases in early 2019 (**Figure 2, c, purple**). There is a large decrease in hospital volume and search query data following the public health lockdown of Wuhan on January 23, 2019 (**Figure 2,** third annotation).

This decrease was seen in both hospital and control sites (**Figure 1**). In Spring 2020, hospital volume begins to trend upward again. In late May 2020 there is a small uptick in Baidu search volume for both “diarrhea” and “cough” in Wuhan.

Discussion

Here we show increased hospital traffic and symptom search data in Wuhan preceded the documented start of the SARS-CoV-2 pandemic in December 2019. While we cannot confirm if the increased volume was directly related to the new virus, our evidence supports other recent work showing that emergence happened before identification at the Huanan Seafood market. These findings also corroborate the hypothesis that the virus emerged naturally in southern China and was potentially already circulating at the time of the Wuhan cluster[7].

In August, we identify a unique increase in searches for diarrhea which was neither seen in previous flu seasons or mirrored in the cough search data. While surprising, this finding lines up with the recent recognition that gastrointestinal (GI) symptoms are a unique feature of COVID-19 disease and may be the chief complaint of a significant proportion of presenting patients [18]. This symptom search increase is then followed by a rise in hospital parking lot traffic in October and November, as well as a rise in searches for cough. While we cannot conclude the reason for this increase, we hypothesize that broad community transmission may have led to more acute cases requiring medical attention, resulting in higher viral loads and worse symptoms[19]. This temporal progression of clinical presentation from mild illness to more severe outcomes has been shown elsewhere[20]. Interestingly, a retrospective study was conducted in Wuhan, China at a hospital designated for the management of patients with COVID-19, which also happens to be represented in our dataset (Wuhan Union Hospital, Wuhan Tongji Medical University)[21]. While respiratory symptoms are common indicators of SARS-CoV-2 infection, this study revealed that a potentially large segment of patients with mainly digestive symptoms, such as diarrhea, may play an important role in community transmission.

The initial rise in GI symptoms may also hint at the missed early signals of COVID-19 in current surveillance systems for respiratory pathogens. The standard definition for influenza-like illness is a combination of fever along with cough and/or sore throat. This narrow definition, which has focused on detection of influenza transmission, would have missed milder cases with a different symptom mix that also could include loss of taste and smell. This finding also hints at the need to broaden surveillance efforts to consider novel pathogens that might display a range of unexpected symptoms. Furthermore, the recent uptick in hospital traffic and search engine query data in May coincides with recent reports of new case clusters in Wuhan[22,23].

The use of satellite data does come with limitations that are amplified in densely populated urban areas. The presence of tall buildings cast shadows that block the view of parking lots, requiring images to be taken at noon local time and exactly overhead. Wuhan experienced a significant amount of cloudy weather during November to February 2019 which along with the consistent smog, created limitations in high quality images that could be harvested. We also experienced challenges in acquiring data from Chinese satellite companies. Finally, prior to the

SARS-CoV-2 outbreak, there are relatively limited archived images of Wuhan compared to other urban centers because of lack of commercial interest.

There are also several limitations to using search query data. We are unable to know the intention of a search and not all symptom searches are necessarily linked to disease morbidity. Search queries resulting from panic and media attention have been shown previously[24] and may have driven the symptom spike we see in January. These data are also vulnerable to fluctuations related to events we might not be aware of and individual search behavior changes over time, which may result in spurious signals[25]. Surveillance using web-query data depends on adequate Internet access and Internet penetration in China can be highly variable. However, by the end of 2017, the internet penetration rate was 70.7% in Wuhan which was 14.9% higher than the national average[26]. Additionally, the use of an automated tool to digitize images does mean the resulting times and values extracted are approximate, although this method has been shown to be effective at replicating time-series[27]. While Baidu query data has been validated for influenza epidemic surveillance[14], investigation of the viability of these data to monitor COVID-19 is still in progress[28]. Nonetheless, when looking at Google search query for the United States, we see a similar pattern of rising GI and cough symptoms alongside confirmed cases.

Although further research is needed to validate the emergence of SARS-CoV-2, this study adds to a growing body of work on the value of digital sources as an early indicator of a disease outbreak in the context of limited integrated electronic surveillance data. While not a replacement for more traditional methods, these data can help supplement other sources to provide a richer situational awareness picture of social disruptions, including those caused by a novel respiratory virus.

Author contributions: EON and JSB contributed to conceptualization. YB, LG and JSB contributed to data acquisition. EON, BR and JSB contributed to data analysis. BR and JSB wrote the first draft of the manuscript. All authors contributed to interpretation of results and final manuscript writing. All authors have seen and approved the manuscript.

Competing interests: The authors did not receive funding for this work and have no conflicts of interest to declare.

References

1. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. 2020;395(10224):565-574. doi:10.1016/S0140-6736(20)30251-8
2. Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new coronavirus epidemic: Evidence for virus evolution. *J Med Virol*. 2020;92(4):455-459. doi:10.1002/jmv.25688
3. Duchene S, Featherstone L, Haritopoulou-Sinanidou M, Rambaut A, Lemey P, Baele G. Temporal signal and the phylodynamic threshold of SARS-CoV-2. *bioRxiv*. May 2020:2020.05.04.077735. doi:10.1101/2020.05.04.077735
4. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497-506. doi:10.1016/S0140-6736(20)30183-5
5. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382(8):727-733. doi:10.1056/NEJMoa2001017
6. Wuhan's Huanan seafood market a victim of COVID-19: CDC director - Global Times. https://www.globaltimes.cn/content/1189506.shtml. Accessed June 1, 2020.
7. Latinne A, Hu B, Olival KJ, et al. Origin and cross-species transmission of bat coronaviruses in China. *bioRxiv*. May 2020:2020.05.31.116061. doi:10.1101/2020.05.31.116061
8. Deslandes A, Berti V, Tandjaoui-Lambotte Y, et al. SARS-COV-2 was already spreading in France in late December 2019. *Int J Antimicrob Agents*. May 2020:106006. doi:10.1016/j.ijantimicag.2020.106006
9. Zhan SH, Deverman BE, Chan YA. SARS-CoV-2 is well adapted for humans. What does this mean for re-emergence? *bioRxiv*. May 2020:2020.05.01.073262. doi:10.1101/2020.05.01.073262
10. Salathé M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS. Influenza A (H7N9) and the importance of digital epidemiology. *N Engl J Med*. 2013;369(5):401-404. doi:10.1056/NEJMp1307752
11. Nsoesie EO, Butler P, Ramakrishnan N, Mekaru SR, Brownstein JS. Monitoring Disease Trends using Hospital Traffic Data from High Resolution Satellite Imagery: A Feasibility Study. *Sci Rep*. 2015;5(1):1-8. doi:10.1038/srep09112
12. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection - Harnessing the web for public health surveillance. *N Engl J Med*. 2009;360(21):2153-2157. doi:10.1056/NEJMp0900702
13. Dai Y, Wang J. Identification of COVID-19 Outbreak Signals Prior to the Traditional Disease Surveillance System. *SSRN Electron J*. May 2020. doi:10.2139/ssrn.3578808
14. Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS. Monitoring Influenza Epidemics in China with Search Query from Baidu. Cowling BJ, ed. *PLoS One*. 2013;8(5):e64323. doi:10.1371/journal.pone.0064323
15. Rohatgi A. WebPlotDigitizer. 2019. https://automeris.io/WebPlotDigitizer/blog/index.html. Accessed June 1, 2020.
16. Kong WH, Li Y, Peng MW, et al. SARS-CoV-2 detection in patients with influenza-like illness. *Nat Microbiol*. 2020;5(5):675-678. doi:10.1038/s41564-020-0713-1
17. Xu B, Gutierrez B, Mekaru S, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci Data*. 2020;7(1):1-6. doi:10.1038/s41597-020-0448-0
18. Pan L, Mu M, Yang P, et al. Clinical Characteristics of COVID-19 Patients With Digestive Symptoms in Hubei, China. *Am J Gastroenterol*. 2020;115(5):766-773. doi:10.14309/ajg.0000000000000620
19. To KKW, Tsang OTY, Leung WS, et al. Temporal profiles of viral load in posterior

oropharyngeal saliva samples and serum antibody responses during infection by SARS-CoV-2: an observational cohort study. *Lancet Infect Dis*. 2020;20(5):565-574. doi:10.1016/S1473-3099(20)30196-1

20. Brownstein JS, Kleinman KP, Mandl KD. Identifying pediatric age groups for influenza vaccination using a real-time regional surveillance system. *Am J Epidemiol*. 2005;162(7):686-693. doi:10.1093/aje/kwi257
21. Han C, Duan C, Zhang S, et al. Digestive Symptoms in COVID-19 Patients With Mild Disease Severity. *Am J Gastroenterol*. April 2020:1. doi:10.14309/ajg.0000000000000664
22. 宣传处. *Hubei Government Situation Report, May 10, 2020*.; 2020. http://wjw.hubei.gov.cn/fbjd/tzgg/202005/t20200511_2266405.shtml.
23. 武汉新增5例确诊患者 来自同一小区-新华网. http://www.xinhuanet.com/politics/2020-05/11/c_1125968733.htm. Accessed June 1, 2020.
24. Butler D. When Google got flu wrong. *Nature*. 2013;494(7436):155-156. doi:10.1038/494155a
25. Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from Google flu trends? *Am J Prev Med*. 2014;47(3):341-347. doi:10.1016/j.amepre.2014.05.020
26. 武汉：全市网民770万 超98%用手机上网_全国. https://www.sohu.com/a/284177554_119038. Accessed June 1, 2020.
27. Nagar R, Yuan Q, Freifeld CC, et al. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *J Med Internet Res*. 2014;16(10):e236. doi:10.2196/jmir.3416
28. Lu T, Reis BY. Internet Search Patterns Reveal Clinical Course of Disease Progression for COVID-19 and Predict Pandemic Spread in 32 Countries. *medRxiv*. 2020:2020.05.01.20087858. doi:10.1101/2020.05.01.20087858

Figures

Figure 1: Parking lot volume in a Wuhan, China hospital and control site

Satellite images and counts of cars (red) and trucks (yellow) in 2 Wuhan, China, parking areas. Top panel is Wuhan Tianyou Hospital pre-epidemic (***A***), October 2019 (***B***) and during the height of the COVID-19 outbreak (***C***). Lower panel is the Huanan Seafood Market, a high-volume control site, in September 2019 (***D***) and February 2020 (***E***).

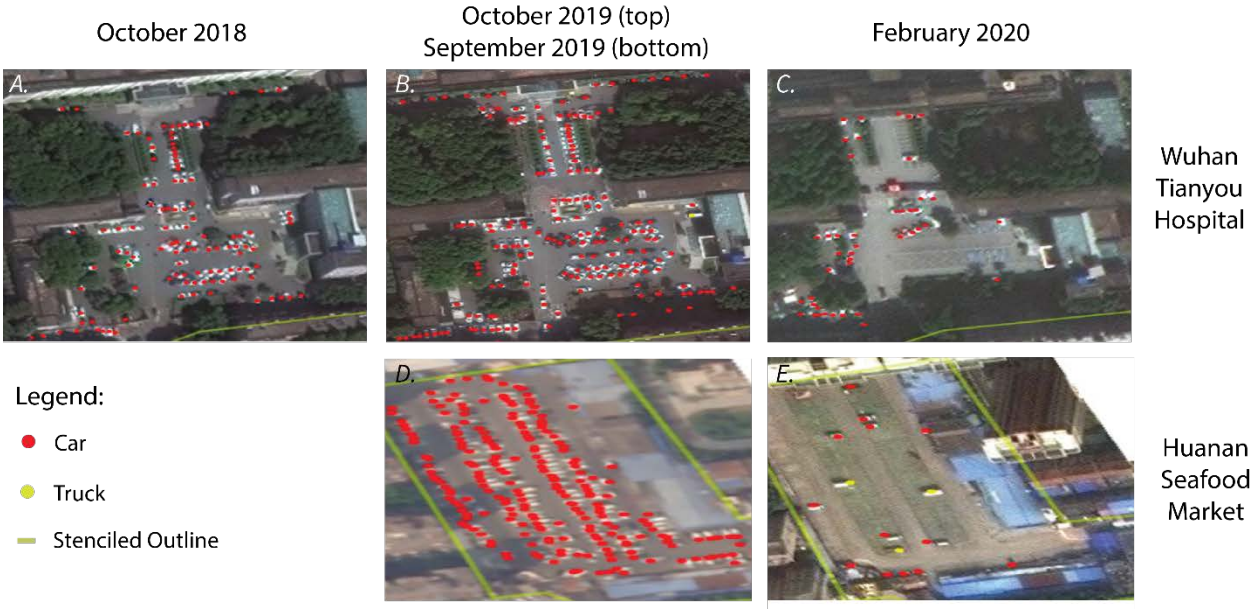

Figure 2: Time-series of different ILI, symptoms and surveillance signals, January 2018 - May 2020

Daily parking lot volume normalized to the same hospital's baseline (weekday, Saturday or Sunday) for 6 Wuhan, China Hospitals (***a***) from January 2018 - April 2020 and a fitted loess regression (***a***, orange) with standard errors. Relative Baidu search query volume is shown for the terms "cough" (***b***, blue) and "diarrhea" (***b***, red) between January 2018 - May 2020. Three yearly influenza-like-illness cycles (***c***, turquoise, right side axis) from two Wuhan hospitals plotted alongside confirmed COVID-19 case counts (***c***, purple, left side axis) extracted from open-source disease data. Date annotations (vertical dashed lines) on all three panels representing August 1, 2019 (rise in traffic and diarrhea signal), December 1, 2019 (date of first confirmed COVID-19 case) and January 23, 2020 (implementation of the Wuhan public health lockdown).