Student Name: Jesseca Johnson

Course: DSC 501: Introduction to Data Science

Week 2 Article Review and Critique

Article Information

The article I have chosen is "Analysis of hospital traffic and search engine data in Wuhan China indicates early disease activity in the Fall of 2019" by Elaine Okanyene Nsoesie, Benjamin Rader, Yiyao L. Barnoon, Lauren Goodwin, and John S. Brownstein, who all contributed equally. The article is nine pages long, including its References section. It was published in 2020 as a Harvard Medical School scholarly article on their DASH website (Digital Access to Scholarship at Harvard). The authors declared no conflict of interest.

Article Summary

This article proposes that although the COVID-19 pandemic is thought to have originally been precipitated by a "zoonotic spillover event" in November or December 2019 at the Huanan Seafood Market in Wuhan, the virus may have been circulating in the community before that time. The authors explore this possibility using two methods of research: first, they examine satellite data and conduct car counting at several large hospitals in Wuhan, finding that car counts increased in late summer/early fall 2019; secondly, using Baidu search engine data from Wuhan, they examine the number of instances of searches for "cough" and "diarrhea" from January 2018 – May 2020 (Nsoesie, Rader, Barnoon, Goodwen & Brownstein, 2020).

Regarding the hospital traffic data, the authors obtained high-resolution satellite imagery for Wuhan using Remote Sensing Metrics (it seems that this company also completed their car-counting). They made a list of Wuhan area hospitals, and after excluding some sub-specialty hospitals and hospitals without satellite imagery available, performed their analysis on six:

1. Hubei Women and Children's

2. Wuhan Central
3. Wuhan Tianyou
4. Wuhan Tongji Medical University
5. Wuhan Union
6. Zhongnan Hospital of Wuhan University

Using a simple formula to find the mean, they fit a LOWESS (locally estimated scatterplot smoothing) line to the data and found that there was an increase in parking lot volume starting in August 2019 (see Figure a) (Nsoesie et al., 2020).

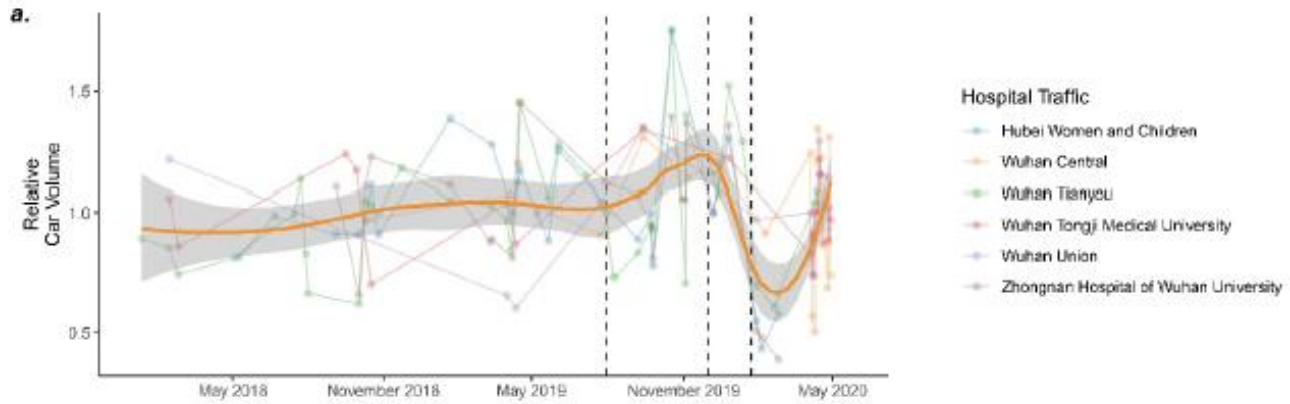

Figure A: Relative car volume of six Wuhan hospitals, January 2018-May 2020 (Nsoesi et al., 2020). The dashed lines represent August 1, 2019, December 1, 2019, and January 23, 2020 (the first day of the Wuhan lockdown), respectively

For the Baidu (search engine) data, the group extracted the relative search volumes of the words “cough” and “diarrhea” using WebPlotDigitizer, v4.2, from April 2017 to May 2020. They found that there was increase in searches for “diarrhea” starting in July 2019 and for “cough” starting in August 2019 (see Figure B) (Nsoesie et al., 2020).

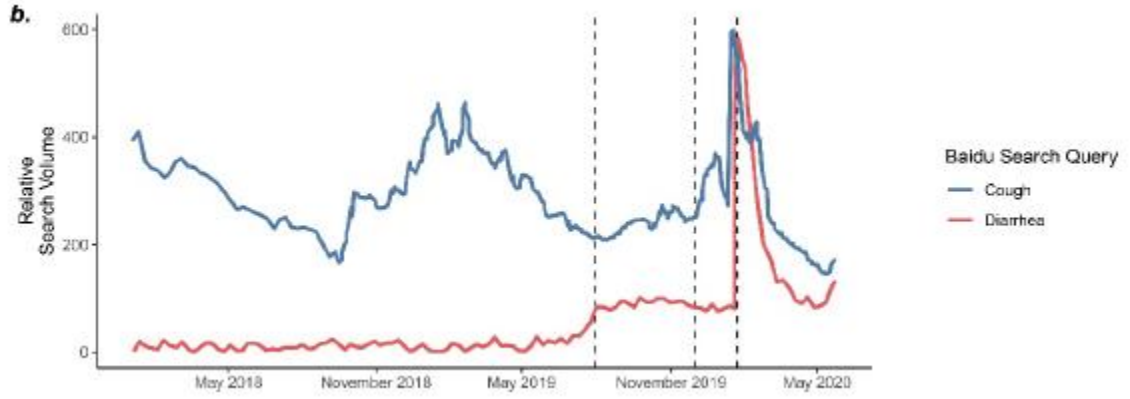

Figure B: Relative Baidu search volume for "cough" (blue line) and "diarrhea" (red line) (Nsoesi et al., 2020). The dashed lines represent August 1, 2019, December 1, 2019, and January 23, 2020 (the first day of the Wuhan lockdown), respectively.

Research Project Relationship

I selected this article because, like many other people around the world, I am interested in the origin of COVID-19 and when exactly it began spreading. The most prominent theory about the origin of the virus is that it came from an animal at a wildlife farm in Yunnan or a nearby province and that that infected animal was transported and sold at Huanan Seafood Market (WHO, 2021). However, the question of exactly *when* this happened remains. Was there community transmission months, or even years, prior to the major spread that started at the end of 2019 and continued into 2020? It is important to understand all the facts to inform future policies on preventing and managing future outbreaks of zoonotic origin. Therefore, I would like to perform my own follow-up study to this study, which would involve looking at search engine data only (not hospital car volume) in two cities: first, Kunming, because it is the capital of and largest city in Yunnan; and second, Guangzhou, because it appears that Guangzhou is the number one fight destination from Wuhan, with 128 flights a week as of May 2021 ("Wuhan Tianhe International Airport," 2021). I would like to add more search terms specific to COVID-19 based on other search terms that increased with rising cases, specifically unique ones like loss of taste or smell (while "cough" is important, it is more difficult to control for that variable because coughing can be a symptom of many other conditions, such as influenza).

Article Critique

Because car counting is a relatively new form of data collection, and because the authors only provided a small selection of the images they used for this purpose, one must entertain some skepticism in how this part of the data collection was performed and whether it was accurate. (The topic of new kinds of data, including images, is touched on by O'Neil and Schutt on page 23 of *Doing Data Science,* though only to say that it isn't covered in the book. It is in fact so new that there is very little literature about it!) The authors also do not address the issue of any other potential causes or correlations of increased hospital car volume in late summer to early fall 2019. For example, was there an outbreak of another disease around that time that could have resulted in increased hospitalizations? It would be a good idea to at least search local newspapers for any indication. Is it possible that there was a large increase in the number of cars circulating in Wuhan in general between 2018 and 2019? Perhaps there is some data about car ownership in China that could help rule out this possibility.

For the search data, as mentioned in the Research Project Relationship section, "cough" and "diarrhea" are but two covid symptoms of many that could be examined. Others could be "fever," "difficulty breathing," "fatigue," "headache," "sore throat," "loss of taste/smell," or "nausea" (*Symptoms of COVID-19*, 2021). Another issue with "cough" is that it is also frequently searched for during flu season. The authors did mention this and it appears that in their study, an increase in searches for "cough" was mostly aligned with increases in influenza-like illness and did not begin increasing in August like "diarrhea" did (thus not lending support to their hypothesis). (The fact that they did disclose this fact and were not trying to pass the increase off as definitely being related to COVID-19 adds legitimacy to their findings, as they could have tried to do so – this technique is described in Wheelan's *Naked Statistics* in Chapter 3 as "deceptive description.")

Conclusion

"Analysis of hospital traffic and search engine data in Wuhan China indicates early disease activity in the Fall of 2019" is important in several ways aside from the obvious implication that COVID-19 spread may have begun earlier than currently understood. First, it attempts to further solidify the legitimacy of using "alternative data," a relatively new form of information collecting, not only in general, but as a window into nations like China that may not always be forthcoming with their data, motivations, or other information. Second, it stands out in its use of ground-level data coming from the general population rather than from health care facilities, governments, or other, more established sources of data; again, the issue of more secretive states such as China and how to get around their imposed obstacles comes to mind. Third, the article also demonstrates that, although we must be careful with the work of amateur data sleuths, it is becoming more feasible each day for the average person, without the benefit of a medical education or access to any type of lab, to conduct their own research – even from the comfort of their own home – and perhaps contribute meaningful knowledge to the medical or any other field.

With so many eyes on the data, information gathering could happen more quickly than it has in the past. O'Neil and Schutt mention this "big data" concept on page 24 of *Doing Data Science* – one definition of big data is that it is a cultural phenomenon, describing just how much data collection and analysis is going on in our daily lives. Likewise, it is now perfectly possible for a regular member of the public to turn this trend into something that, rather than something that others use to try to capitalize on *them* and their interests, movements, etc., satisfies *their* own curiosity using open data and data science methods about any given subject, thus adding another dimension to this definition and transforming them from a passive to an active actor in the big data machine. This is just as applicable to COVID-19 as to anything else; there are still many mysteries within the pandemic and these new data collection methods could perhaps be used to shed some light on some of them.

Works Cited

Nsoesie, E., Rader, B., Barnoon, Y., Goodwin, L., and Brownstein, J. (2020). Analysis of hospital traffic and search engine data in Wuhan China indicates early disease activity in the Fall of 2019. *Harvard Medical School (HMS) Scholarly Articles*. https://dash.harvard.edu/bitstream/handle/1/42669767/Satellite_Images_Baidu_COVID19_manuscript_DASH.pdf?sequence=3&isAllowed=y

O'Neil, C. and Schutt, R. (2014). *Doing Data Science*. O'Reilly Media, Inc.

Symptoms of COVID-19. (2021, February 22). Centers for Disease Control and Prevention (CDC). Retrieved May 22, 2021 from https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

Wheelan, C. (2013). *Naked Statistics: Stripping the Dread from the Data*. W. W. Norton & Company, Inc.

World Health Organization (WHO). (2021). *WHO-convened Global Study of Origins of SARS-CoV-2: China Part (Joint WHO-China Study)*. WHO. https://www.who.int/docs/default-source/coronaviruse/final-joint-report_origins-studies-6-april-201.pdf?sfvrsn=4f5e5196_1&download=true

Wuhan Tianhe International Airport. (2021, May 22). Flightradar24. Retrieved May 22, 2021 from https://www.flightradar24.com/data/airports/wuh