Examining the Correlation of Google Influenza Trend with Hospital Data: Retrospective Study

Abdulrahman M Jabour 1
Joe Varghese 1
Ahmed H Damad 2
Khalid Y Ghailan 3
Asim M Mehmood 1

1 Health Informatics Department, Faculty of Public Health and Tropical Medicine, Jazan University, Jazan, Saudi Arabia; 2 Quality & Patient Safety Department, King Fahd Central Hospital - Jazan, Jazan, Saudi Arabia; 3 Epidemiology Department, Faculty of Public Health and Tropical Medicine, Jazan University, Jazan, Saudi Arabia

Introduction: Many studies have explored social media and user search activities such as Google Trends to predict and detect influenza activities. Studies that examined Google Trends correlation with the actual hospital influenza cases were conducted in non-tropical regions that have clearly defined seasons. Tropical areas are known for having less-defined seasonality and the extent of Google Trends concordance with actual influenza cases is unknown for these areas. The goal of this study is to compare Google Trends with hospital cases in tropical regions.

Methods: We analyzed 48,263 influenza cases in the time period of 2010 to 2019. The cases were retrieved from central hospital medical records in tropical regions using the corresponding codes for influenza ICD-10 AM. Cases from the medical records were compared with Google Trends to determine trends, seasonality, and correlation.

Results: Graphically, there were some similar areas of the trend, but cross-correlation analysis did not show any significant correlation between hospital and Google Trends with a maximum correlation rate of 0.300. Seasonality analysis showed a clear pattern that peaked around November in Google Trends while hospital data showed less defined seasonality with a smaller peak occurring at the end of December and beginning of January.

Conclusion: Based on the results, there is a weak correlation between Google Trends and hospital data. More innovative methods are emerging to predict influenza activity using social media and user search data and further study is needed to examine the concurrent trends derived using these methods across regions that have different humidity levels and temperatures.

Keywords: influenza, predictive, surveillance, data quality, disease outbreaks, big data, epidemiological monitoring

Background
Seasonal influenza is an acute respiratory infection caused by the influenza virus. There are various forms of seasonal influenza and it spreads easily and rapidly in communities throughout the world. The virus is able to assume epidemic proportion in a short time frame.1 In temperate climates, seasonal epidemics occur primarily during the winter, but in tropical regions, influenza may occur throughout the year, something that causes irregular and hard to predict outbreaks.2,3

Surveillance systems for influenza often rely on hospital-based clinical data to monitor the virus’ activities and allow public health officials to take necessary actions.4 Different surveillance systems may have different criteria and reporting eligibilities. The determination of data included in this report is determined by the surveillance system. The assessment of case eligibility for reporting requires data
from multiple sources including physical examination, collection of signs and symptoms, and laboratory results.

The lifecycle of an influenza outbreak has several phases and the detection in each phase requires different information. The lifecycle begins with initial exposure and then progresses to incubation, symptoms, a visit to a healthcare facility, and confirmation of diagnosis. There are various surveillance systems that can be used at different phases of the disease lifecycle. Each system has its own strengths and weaknesses that can be identified based on the data being collected. Generally, systems can be in one or two categories: data collected before visiting a healthcare facility and data collected after visiting a healthcare facility. Syndromic surveillance methods involve “relying on detection of clinical case features that are discernible before confirmed diagnoses are made.” This type of system often uses data prior to confirmation of diagnosis. It is an early stage of detection that uses information to estimate the probability of a case or outbreak and then may ask for a further public health response. This information occurs prior to diagnosis, something which often requires laboratory results to be obtained.

Some systems require laboratory confirmation of diagnosis. Although this type of system can improve the accuracy of the reported data by allowing physicians to assess the case and confirm the diagnosis, it has the disadvantage of longer reporting time as well as the possibility of missing false-negative cases. This system could result in longer outbreak detection time since there are additional steps needed prior to the data being entered into the reporting system. The time taken for analysis at the central registry and the time needed for the public health registry to communicate the findings with the applicable hospital(s) authority and treating physician(s) and the issuing of a warning signal can be somewhat substantial.

Another limitation of the system is that some physicians may miss cases. This issue is more probably with Influenza Like Illness (ILI) as it shares similar signs and symptoms with other diseases, especially during the early stages of the infection. Confirmation is typically obtained through laboratory testing. Physician awareness of an outbreak in an area will increase the likelihood of ordering laboratory resting and engaging in a more thorough investigation. Early warning signs of unusual influenza activity in an area can direct physician attention toward a potential outbreak and greater investigation. This added attention can ultimately lead to a proper diagnosis of the epidemic and the triggering of necessary public health actions.

With the advancement of business intelligence and the availability of “big data” sources that provide information collected from multiple sources, researchers have started to explore the value of using non-hospital data for outbreak detection. One of the common sources of data is Google search terms, also known as Google Trends. Google Trends can be used to analyze search keywords, in this case searches for specific diseases. Unlike hospital data, Google Trends can provide early warning signals as it collects and analyses online activity prior to patients visiting healthcare facilities. Another advantage of this system is that Google Trend collects additional information that may not be within the data set collected by hospitals or reported to public health agencies.

Despite the popularity of the Google influenza trend, studies in different countries have provided conflicting results of how Google Trend data reflect the actual trends of influenza based on hospital cases. These studies have indicated that more research is needed if Google Trends will be used for forecasting purposes. Studies about how Google Trends correlates with actual cases during off-peak periods and smaller waves are also needed. While many studies indicated the value of Google Trends for complementing the traditional methods, others pointed out some limitations related to the reliability and accuracy of Google Trends when compared with seasonal or pandemic influenza.

Prior studies were conducted in climates with clearly defined seasonality, but areas with more tropical climates are known for having smaller peaks and less-defined influenza seasonality. The extent to which Google Trends reflects actual influenza activity in a tropical region or hospital level is unknown.

Given the documented impact of climate on the trend of influenza cases, little is known if Google Trends will be impacted by climate condition to the same extent. More studies about the correlation between of influenza cases and Google Trends at different climate will improve our understanding of the limitation and strength of Google Trends.

Determining the accuracy of Google Trend data in different regions can unlock a wide range of public health and epidemiological applications. In this study, we will compare influenza-related online activity based on Google Trend with actual influenza cases in a hospital.
Methods
In this study, we compared influenza activity trends on Google Trend with the actual trend based on hospital influenza cases. We created time-series and seasonal graphs. This study was approved by the institutional review board (IRB) at the Saudi Ministry of Health, Jazan Health Affairs, protocol number H-10-Z-073. The study conducted in compliance with the Declaration of Helsinki and the ethical guidelines of ensuring that every precaution has been taken to protect the privacy of research subjects and the confidentiality of their personal information. All data employed in the study were de-identified secondary data and no patients identifying information were included. Therefore, explicit patients consent to review their medical records at this case was not required by the Saudi Ministry of Health.

Google Trend Data
We carried out a search in Google Trend (https://trends.google.com) using the keywords “influenza”, “corona”, and “MERS” for a 10-year period starting in January 2010 and ending in December 2019. The search trends pertained to the entire geographical territory of the Kingdom of Saudi Arabia. Google Trends uses the proportion of cases that peaked during a specified period and the data was graphed along with data from hospital record systems. The collection method and use of Google Trends data complied with the terms and conditions of the website (https://policies.google.com/terms?hl=en-US).

Hospital Data
The influenza incidence data used in this analysis were retrieved from the Electronic Health Records (EHRs) of King Fahad Central Hospital which is tertiary-level care hospital. It is a government hospital that operate under the Saudi Ministry of Health. Therefore, the IRB for the study was obtained from the Saudi Ministry of Health. The hospital is in Jazan, by the Red Sea, in the southwestern part of Saudi Arabia. This region has a somewhat tropical climate with an average temperature range of 78°F to 91°F and an average humidity of over 64% (more about the weather can be found in Table 1).

We used the ICD-10AM codes for Influenza and Influenza Like Illness (ILI) spanning from 2010 to 2019. All flu cases retrieved for the study were identified based on the direct influenza codes (J09, J10, J10.1, J10.81, J10.89, J11, J11.1, J11.8) and the ILI codes identified in prior studies. We also mapped to the earlier version of ICD for the period before the ICD-10AM implementation.

We did not use the criteria of fever and cough as indicators for the viruses under study since they are not specific to ILI like the other codes used in this study area. All data retrieved were anonymized before being accessed by the research team and no personal identifying information was collected.

Analysis
All data are entered in the hospital EHR included a unique medical records number and visit ID. Because this study is interested in the number of events, the codes were aggregated based on the number of visits. Patients who may have multiple diagnoses were counted as a single data point and anonymized before being shared for research. The annual number of hospital incidents retrieved is displayed in Table 2.

For comparison, both Google Trends and hospital data were converted based on the two variables of time and intensity. The intensity was calculated as the monthly proportion of incidents from its peak within the timeframe of 2010 through 2019. Data were graphed together to compare the alignment and patterns of change.

The pattern of the influenza season and peaks were examined by plotting the seasonal graph for each data source. The annual trends for each year were recorded separately to compare changes in the pattern of influenza over time and at each data source. The Statistical Package for the Social Sciences (SPSS) v 21 (IBM Corp) was used for seasonality analysis and visualization. We also conducted a cross-correlation analysis.
to examine the association between the trends generated by Google and hospital cases. In addition, cross correlation enables us to examine if an association exist with the lag (delay) between the two time series trends (Google and hospital cases). We used the same time frame used for the seasonal and trend analysis (2010–2019) and the same unit of time interval (months). The cross-correlation results were interpreted as follow; weak (<0.4) Moderate (0.4–0.6), strong (>0.6).26

Results
Using hospital medical records, 48,263 influenza cases spanning from 2010 to 2019 were retrieved. These cases were compared with influenza activities found by Google Trends for the same time period. The results of our study will include a time-series trend comparison, seasonality analysis, and cross-correlation analysis.

Trends and Time-Series Analysis
The comparison of Google Trend results (Figure 1) with hospital influenza incidence patterns also included the Google search for the terms “MERS”, and “corona”. Of the three keywords, only “influenza” showed a noticeable level of online activity that might correlate with the hospital trend. The only exception is a high online activity for “MERS” during 2014. This activity corresponds with the onset of this infection in the country.27

Google Trends showed a much lower volume in the initial periods compared to the second half of the study period. These results could reflect the lower use of internet searches in the period prior to 2015. For the search of “influenza”, a noticeable overlap can be observed between Google Trend results and hospital data starting from 2015 and ending in December 2019. Through the seasonal peaks of the hospital data trend, there has been a corresponding rise in the result of the keyword “influenza” based on Google Trend data. The hospital data also showed remarkable differences in the number of peaks from year to year.

From 2010 to 2014, Google Trends and hospital cases showed somewhat similar pattern with different magnitudes. Both trends and hospital cases peaked during December 2010, December 2011, and November 2012. They had slightly similar moves during 2013 and 2014. The difference in magnitude can be explained based on how the proportion of cases was calculated in relation to the peak for the entire study period.

Figure 1 also shows that there is a lag in peaking time between Google Trends and hospital data from 2015 to 2017. We analyzed the overlap of disease occurred during the two seasonal peaks between the last quarter of 2015 through the first quarter of 2017 and compared the peaks

Table 2 The Number of Annual Cases Retrieved from the Hospital Database from 2020 to 2019

| Year | Number of cases |
|------|-----------------|
| 2010 | 2025            |
| 2011 | 4430            |
| 2012 | 6780            |
| 2013 | 5994            |
| 2014 | 5183            |
| 2015 | 5981            |
| 2016 | 5695            |
| 2017 | 5054            |
| 2018 | 4389            |
| 2019 | 2732            |
| Total| 48,263          |

Figure 1 Comparing the patterns of the monthly influenza activities from 2010 to 2019 using hospital data, and Google Trends with the terms “influenza”, “MERS”, and “corona”.
to the corresponding results for “influenza” in Google Trend. Figure 1 shows that Google Trends showed increased levels of online searches for “influenza” before the disease peaks of May 2015, February 2016, and January 2017. There is also an overlapping peak in November 2017.

**Seasonality Analysis**

The Google Trend data for the term “influenza” show clear seasonal disease trends over the 10-year period (Figure 2). The onset of disease peak starts in the last quarter of the year and subsides in the first quarter of the following year. There are variations in disease incidence intensity and the exact time of onset each year. The hospital data showed much less defined seasonality (Figure 3) during the last 3 years of the examined period (2017 to 2019). From 2010 to 2016, the peak season starts in October and peaks by the end of the year. It begins to gradually decline around the middle of the year (June).

**Cross-Correlation Analysis**

To examine the correlation between Google Trends and hospital data, we conducted a cross-correlation analysis (Figure 4). In this figure, the cross-correlation between the two trends was generally weak. The highest correlation was observed at lag 0 and lag 5 with 0.299 and 0.300, respectively.

**Discussion**

Assessing how Google Trend represents actual influenza cases opens a wide range of practical epidemiological and public health applications. Studies documented the impact of influenza on hospitalization and how surveillance can inform decision making. Surveillance from search data like Google Trends has the potential of being an early indicator and enable the timely preparedness and intervention.

Several studies have examined the relationship between Google Trends and hospital influenza cases, but the hospitals used in these studies were in non-tropical areas and had clearly-defined seasonality. While previous studies reported conflicting results about Google Trend data reliability, no studies were found to examine how Google Trend data can be used to confirm actual influenza cases in different weather conditions. This study compares hospital data with Google Trends data of 10 years to explore the potential usefulness of Google Trends in surveillance and prediction of influenza in tropical regions.

Our analysis shows a weak relationship between hospital influenza cases and Google Trends. There was some similarity between the two during peak influenza season from 2010 to 2015. There was approximately a month lag between the two during the peak influenza season for 3
Figure 3: The monthly proportion of influenza activities from the hospital data 2010–2019.

Figure 4: A cross-correlation analysis between Google Trends and hospital data for influenza cases.
years. The time-series analysis shows some similarities in certain years and the cross-correlation analysis shows weak correlation between the two trends.

The application of internet data for health research is an evolving field and the domain needs further refinement to generate actionable intelligence for disease control efforts. A study of Google Flu Trends in the United States provided timely and accurate estimates of the influenza activity, especially during epidemic peaks, however, the study also reported that more research is needed to explain the Google Trends accuracy during smaller waves and off-peak periods. Our analysis shows a similar experience with smaller disease peaks. In a study that assessed the sensitivity of search data on Google, Twitter, and Wikipedia, there was low to medium sensitivity in influenza event detection. The study concluded that the search data had a limited ability to contribute to surveillance systems designed to detect influenza. Similarly, cross-correlation analysis in our study showed a weak correlation. Influenza patterns in tropical regions do not always follow seasonal patterns and there is high background influenza activity throughout the year.

Despite its benefits, Google Trends is prone to bias and can provide inaccurate signals. For example, peaks could be overestimated when users search of post-influenza-related keywords based on media coverage. In these cases, individuals who do not suffer from influenza-related symptoms would impact the results. According to Cervellin et al., Google's Trends is more influenced by media coverage than the actual epidemiological burden. The spike in the use of the search term “MERS” during 2014 was the result of high public anxiety due to media coverage and was not an accurate reflection of the cases during this time.

Limitations

One of the limitations was the difference in the level of geographical coverage of Google Trends and hospital data. In our study, the number of cases generated by Google Trends at a city or regional level was low and made inferences difficult. We used the country-level Google Trend for this reason. The extent to which Google Trend search terms represent a particular geographic region should be taken into account. This is especially important when a country or state has areas with different environmental factors, in particular humidity and temperature. The data collection area could have caused some cities and areas with higher populations to be overrepresented. Google Trends data for the study’s hospital is available in smaller quantities, a factor that makes it somewhat difficult to use.

Conclusions

Our study showed a weak correlation between Google Trends and hospital influenza data. Although the two trends had some similarities during the peak influenza season, the overall correlation was statistically weak. In general, influenza cases in hot and humid areas have less defined seasonality compared to non-tropical areas. Given the strong association between weather and influenza. We found it very important to examine the accuracy of Google Trends of influenza under the different regions with different climates and the degree to which regional climates can impact the accuracy of Google Trends. Further studies could also examine the alternative sources of influenza data derived from search terms and social media. The author also recommends examining ILI compared to pandemic influenza as the search term for “pandemic influenza” may be more a product of media coverage than the actual virus.

Abbreviations

HER, Electronic Health Records; ICD-10 AM, The International Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification; ILI, Influenza like Illness; MERS, Middle East Respiratory Syndrome.

Acknowledgments

The authors would like to thank Mr Asim Mahmood for his effort in the data retrieval.

Author Contributions

All authors contributed to data analysis, drafting or revising the article, gave final approval of the version to be published, agreed to the submitted journal, and agree to be accountable for all aspects of the work.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Disclosure

The authors declare that they have no competing interests.
References

1. Iuliano AD, Roguski KM, Chang HH, et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. Lancet. 2018;391(10127):1285–1300. doi:10.1016/S0140-6736(17)33293-2

2. Tamerius JD, Shaman J, Alonso WJ, et al. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. PLoS Pathog. 2013;9(3). doi:10.1371/journal.ppat.1003408

3. Barreca AI, Shimshack JP. Absolute humidity, temperature, and influenza mortality: 30 years of county-level evidence from the United States. Am J Epidemiol. 2012;176(supp1):S114–S122. doi:10.1093/aje/kws259

4. M'Ikanatha NM, Lynfield R, Van Beneden CA, de Valk H. Influenza surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc. 2004;11(2):141–150. doi:10.1197/jamia.M1135

5. Mandl KD, Overhage JM, Wagner MM, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc. 2004;11(2):141–150. doi:10.1197/jamia.M1135

6. Lin PL, Lee TH, Rowe EK. Middle East respiratory syndrome coronavirus (MERS CoV): update 2013. Curr Infect Dis Rep. 2013;15(4):295–298. doi:10.1007/s11908-013-0344-2

7. Alraddadi BM, Watson JT, Almarashi A, et al. Risk factors for primary middle east respiratory syndrome coronavirus illness in humans, Saudi Arabia, 2014. Emerg Infect Dis. 2016;22(1):49–55. doi:10.3201/cid.220115340

8. Conzade R, Grant R, Malik MR, et al. Reported direct and indirect contact with dromedary camels among laboratory-confirmed MERS-CoV cases. Viruses. 2018;10(8):425. doi:10.3390/v10080425

9. ECDC. Epidemiological update: Middle East respiratory syndrome coronavirus (MERS-CoV). European centre for disease prevention and control; An agency of the European Union; 2014;

10. Pollett S, Wood N, Boscardin WJ, et al. Validating the use of Google Trends to enhance pertussis surveillance in California. PLoS Curr. 2015;7currents.outbreaks.7119696b3e7523faa4543faa8c7c56e2. doi:10.1371/currents.outbreaks.7119696b3e7523faa4543faa8c7c56e2

11. Jabour AM, Dixon BE. Monitoring public health reporting: data tracking in cancer registries. Online J Public Health Inform. 2018;10(3). doi:10.5210/ojphi.v10i3.9432

12. Jabour AM, Dixon BE, Jones JF, Hagstrom DA. Toward timely data for cancer research: assessment and reengineering of the cancer reporting process. JMIR Cancer. 2018;4(1):e4. doi:10.2196/cancer.7515

13. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PLoS One. 2011;6(8):e23610. doi:10.1371/journal.pone.0023610

14. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol. 2013;9(10):e1003256. doi:10.1371/journal.pcbi.1003256

15. Cho S, Sohn CH, Jo MW, et al. Correlation between national influenza surveillance data and Google Trends in South Korea. PLoS One. 2013;8(12):e81422. doi:10.1371/journal.pone.0081422

16. Sharpe JD, Hopkins RS, Cook RL, Striley CW. Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis. JMIR Public Health Surveill. 2016;2(2):e161. doi:10.2196/publichealth.5901

17. Klembczyk JJ, Jalalpour M, Levin S, et al. Google flu trends spatial variability validated against emergency department influenza-related visits. J Med Internet Res. 2016;18(6):e175. doi:10.2196/jmir.5585

18. Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for influenza surveillance in South China. PLoS One. 2013;8(1):e55205. doi:10.1371/journal.pone.0055205

19. Dugas AF, Jalalpour M, Gel Y, et al. Influenza forecasting with Google flu trends. PLoS One. 2013;8(2):e56176. doi:10.1371/journal.pone.0056176

20. Butler D. When Google got flu wrong; US outbreak foreshadows a leading web-based method for tracking seasonal flu. Secondary when Google got flu wrong; US outbreak foreshadows a leading web-based method for tracking seasonal flu; 2013. Available from: https://www.wired.com/2013/01/when-google-got-flu-wrong-1.12413. Accessed October 18, 2021.

21. Laizer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science. 2014;343(6176):1203–1205. doi:10.1126/science.1248506

22. World Medical Association. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. Bull World Health Organ. 2001;79(4):373.

23. Climate-Data.Org. CLIMATE DATA. Secondary CLIMATE DATA; 2021. Available from: https://en.climate-data.org/asia/saudi-arabia/jazan-region/jizan-3568/. Accessed October 18, 2021.

24. Fitzner J, Qasmieh S, Mounts AW, et al. Revision of clinical case definitions: influenza-like illness and severe acute respiratory infection. Bull World Health Organ. 2018;96(2):122. doi:10.2471/BLT.17.194514

25. Schindeler SK, Muscatello DJ, Ferson MJ, Rogers KD, Grant P, Churches T. Evaluation of alternative respiratory syndromes for specific syndromic surveillance of influenza and respiratory syncytial virus: a time series analysis. BMC Infect Dis. 2009;9(1):190. doi:10.1186/1471-2334-9-190

26. Akoglu H. User’s guide to correlation coefficients. Turkish J Emerg Med. 2018;18(3):91–93. doi:10.1016/j.tjem.2018.08.001

27. World Health Organization. WHO MERS-CoV global summary and assessment of risk. Geneva: WHO; 2017.

28. Bertamino E, Zerbetto A, Capalbo C, et al. Severe Acute Respiratory Syndrome-Corona Virus (MERS-CoV) associated stress among medical students at a university teaching hospital in Saudi Arabia. J Infect Public Health. 2021;14(3):282–287.

29. Wang X, Li Y, O’Brien KL, et al. Global burden of respiratory infections associated with seasonal influenza in children under 5 years in 2018: a systematic review and modelling study. Lancet Global Health. 2020;8(4):e497–e510. doi:10.1016/S2214-109X(19)30545-5

30. Effenberger M, Kronbichler A, Shin JY, Mayer G, Tilg H, Perco P. Association of the COVID-19 pandemic with internet search volumes: a Google Trends™ analysis. Int J Infect Dis. 2020;95:192–197. doi:10.1016/j.ijid.2020.04.033

31. Viboud C, Alonso WJ, Simonsen L. Influenza in tropical regions. PLoS Med. 2006;3(4):e89. doi:10.1371/journal.pmed.0030089

32. Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) Google Flu Trends? Am J Prev Med. 2014;47(3):341–347. doi:10.1016/j.amepre.2014.05.020

33. Cervillen G, Comelli I, Lippi G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. J Epidemiol Glob Health. 2017;7(3):185–189. doi:10.1016/j.jegh.2017.06.001

34. Al-Rabiaah A, Tenshah MH, AL-Eyadhy AA, et al. Middle East Respiratory Syndrome-Corona Virus (MERS-CoV) associated stress among medical students at a university teaching hospital in Saudi Arabia. J Infect Public Health. 2020;13(5):687–691. doi:10.1016/j.jiph.2020.01.005
