PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Google Trends-based non-English-language query data and epidemic diseases: a cross-sectional study of the popular search behavior in Taiwan
AUTHORS	Chang, Yu-Wei; Chiang, Wei-Lun; Wang, Wen-Hung; Lin, Chun-Yu; Hung, Ling-Chien; Tsai, Yi-Chang; Suen, Jau-Ling; Chen, Y.-H.

VERSION 1 – REVIEW

REVIEWER	Vincenza Gianfredi University of Perugia
REVIEW RETURNED	24-Sep-2019

GENERAL COMMENTS	the manuscript is interesting and well written. It is not clear why the Authors chose that time period for the analysis. Please add information on this aspect. There are no acknowledgment of the possibile intrinsic limitation regarding the use of big data on epidemic disease surveillance. For more details I suggest to refer to the following publications: -Monitoring public interest toward pertussis outbreaks: an extensive Google Trends-based analysis -Harnessing Big Data for Communicable Tropical and Sub-Tropical Disorders: Implications From a Systematic Review of the Literature.

REVIEWER	Aaron Secrest University of Utah, United States
REVIEW RETURNED	14-Oct-2019

GENERAL COMMENTS	Overview This is a prospective, observational study comparing public epidemic data from Taiwan to google trends search frequencies to see if google trends can serve as a surrogate surveillance system in a non-English country. Key Points - While the English writing in the paper is well done, there are many unusual phrases and inaccurate wording (e.g., “the excellent surveillance tools” and “even they do not visit hospitals” in the abstract) that show the paper would benefit from having a native English speaker review and edit the paper throughout. Results - Consider rounding your correlation numbers to the hundredths place – I’m not sure the utility of the r-values to the thousandths place – 0.796 and 0.80 mean the same thing to the reader.

	- Not sure if data from Taiwan can be generalized to all non-English-speaking countries. The authors repeatedly state that their findings are for non-English-speaking countries, and I think that statement is too broad. Abstract - Line 50 – “assess” not “access” Results - Page 7, line 14 – I think you mean “no forward”, not “on forward”, and “no forward” sounds weird, consider rephrasing. Discussion - The authors need to discuss why some flu-related symptoms (fever, cough) are highly correlated with flu-like illnesses, but not other symptoms (runny nose)? It seems like the lay-person would be searching all of these. - The authors make this statement: “The web user’s education level, economic situation, cultural and language backgrounds can influence the local habits of Internet searchers.” But also need to clarify that Google Trends does not allow for any capture of demographic information of the Google users. This is a big limitation of google trends. - Page 11, Like 28 – In the US, it is “Twitter”, not “Tweets” for the social media site. Figures 1-4 - These could easily be combined into a 4-panel single figure. - Seeing the temporal association is nice, but it’d be useful to use other health-related search terms that are not related to influenza (like myocardial infarction or diabetes, that lack seasonality) to show discriminant ability of Google Trends. Perhaps all health-related search terms increase around week 15-17 of your study period.

REVIEWER	Jianhua Liu The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China
REVIEW RETURNED	08-Jan-2020

GENERAL COMMENTS	There could be a lot of bias for the model in this study. Personally, I do not agree that by analyzing only the google trend could represent the epidemic outbreaks. For instance, if there is a hot topic recently about the malaria in Africa, the trend of searching related topic may increase in Taiwan as well. But we cannot conclude the prevalence of disease increase in Taiwan base on this information. Moreover, those confounders like education level and age were not included as confounders in this study, which should have both influence for outcome and expose. Therefore, I cannot interpret the result that google trend represent the epidemic disease base on this study.

REVIEWER	Samuli Pesälä University of Helsinki, Helsinki, Finland
REVIEW RETURNED	19-Jan-2020

GENERAL COMMENTS	General comment This study describes the non-English searches from Google Trends to assess epidemic diseases and public opinion through

	popular search behavior. The study compares epidemic data on ILI and EN71 to query data from Google Trends. The results suggest that Google Trends serve as a good surveillance tool for epidemic outbreaks in non-English countries. The study has public health impact in terms of applying traditional register-based data and Google Trends on infectious diseases (ILI, EN71) in order to enhance disease surveillance in the era of Internet. Yet the manuscript needs major revisions. Major comments Introduction: Please include more information on enterovirus infection, its seasonality and epidemics in the Introduction section. Also, a short introduction is needed explaining the seasonality of influenza. Page 4, lines 11-13: Was there a reason why you collected data on influenza in 10/2015-4/2016, but enterovirus in 1-12/2012? Data availability? Please shortly mention the reason. Methods, Page 4, lines 29-31: I suppose that enterovirus analysis included a laboratory test which tested enterovirus infection among patients, or was enterovirus infection based on clinical symptoms? Please clarify. Results, Page 5, lines 37-38: “The total of 8 queries related to influenza (Table 1)”. There are 10 query terms in the table. Which terms related? Page 7, lines 14-15: “Thus, the group C keywords research relative intensities”. What and where is group C? This is not mentioned anywhere else in the manuscript. Figures 1-4, Pages 16-19: Weeks are reported “relatively” (weeks 1-25, i.e. 6 months), not real weeks. Now it seems that an influenza epidemic occurs during weeks 15-23 (April-June). According to other research, influenza weeks in Taiwan should occur something like between weeks 50-14. These should be revised into real weeks. However, in Figure 5, during the year 2012 (weeks 1-52), the peak week (wk 21-26) of an enterovirus epidemic seems to be in the summertime. This seasonality should be explained in the Introduction. Figure 4, Page 19: There is “death of pneumonia and ILI patients” visualized with red vertical bars. This is a little bit unclear. Please choose plural (“deaths”) as shown in Table 2 (“weekly deaths from pneumonia and ILI”). Had these patients had ILI diagnoses before they died of pneumonia? In public opinion estimation, why did you choose the general public’s query term “ECMO” to be compared with positive influenza tests? Please explain. Page 22: Please omit the results from your future studies (Figure in Multimedia Appendix 2, Portuguese keywords). Minor comments In the Abstract (strengths and limitations), “EN71” should be written out when mentioned for the first time in the text, or replaced with “enterovirus infection”.

	In the Abstract, strengths and limitations: “Apart from Google trends, we need to combine more social media to comprehensively analysis the epidemic information through web-related behaviours”. This sentence should be clarified. Results: Page 6, lines 33-34: There is background information on enterovirus (“Enterovirus 71 (EN71) was first identified in California, USA, in 1969.”). I think this should be included in the Introduction section, not Results section. Figure 1 title, Page 16: “…keywords research relative intensity…” Is this a defined/formal term or do you mean “keywords search relative intensity”? In Figure 1, there is “Google Trends search relative intensity” on the right hand side. Page 7, lines 17-18: “…was 4-weeks log between the Internet query data and epidemic advance.” Do you mean that there was a 4-week delay in queries? Then, I think it is a “lag”. Table 3: Page 9, lines 10-11: A typo “entervirus”, should be “enterovirus” Figure 5, Page 20: Two terms are used “enteroviruses infection” and “enterovirus infection”. Please choose “enterovirus infection”. Supplementary material, Page 12: In the table, a typo “ECOM”, should be “ECMO”. Please check the initial letters regarding “Figure” and “Table” and make sure that these are consistently uppercase throughout the text if you refer to the Figures and Tables where you show the data. Also, please check that “Internet” is uppercased in the text.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name

Vincenza Gianfredi

Institution and Country

University of Perugia

Please state any competing interests or state ‘None declared’:

none declared

Please leave your comments for the authors below

1. The manuscript is interesting and well written.
   [Response] We appreciate your kindly comments on our paper.

2. It is not clear why the Authors chose that time period for the analysis. Please add information on this aspect.
   [Response] Thanks for the comments. In the present study, we focused on using Google Trends to monitor the epidemic disease and the public opinion in Taiwan. During this interval (from October 4, 2015, to April 2, 2016), an influenza outbreak, a natural disaster, and an earthquake, were occurred in Taiwan. Thus, the period was suitable to be the research target, which we estimated whether the Google Trends was suitable to be the surveillance tool of the epidemic disease and the public opinion. Moreover, we have evaluated the correlation between Google Trends and the enterovirus infection from 2012 to 2016. The keyword “腸病毒 (enterovirus)” has been shown significant correlation coefficient values with enterovirus infection in these five consecutive years. However, the serious enterovirus outbreak occurred in 2012, which we chose to present this result in the paper. Finally, we have added the information described as above in the text.

3. There are no acknowledgment of the possible intrinsic limitation regarding the use of big data on epidemic disease surveillance. For more details I suggest to refer to the following publications:
   -Monitoring public interest toward pertussis outbreaks: an extensive Google Trends-based analysis
   -Harnessing Big Data for Communicable Tropical and Sub-Tropical Disorders: Implications From a Systematic Review of the Literature.
   [Response] We sincerely appreciate this useful comment. These papers provide us more benefits and aspects to discuss our paper. As the reviewer mention, there are some possible intrinsic limitations in the studies based on the big data for the epidemic observation. We explained and referred these articles in our paper in the introduction and discussion sections.

Reviewer: 2

Reviewer Name

Aaron Secrest

Institution and Country

University of Utah, United States

Please state any competing interests or state ‘None declared’:

None declared

Please leave your comments for the authors below

Overview

This is a prospective, observational study comparing public epidemic data from Taiwan to google trends search frequencies to see if google trends can serve as a surrogate surveillance system in a non-English country.

[Response] Thanks for the kindly suggestions and comments for our work.

Key Points

1. While the English writing in the paper is well done, there are many unusual phrases and inaccurate wording (e.g., “the excellent surveillance tools” and “even they do not visit hospitals” in the abstract) that show the paper would benefit from having a native English speaker review and edit the paper throughout.
   [Response] This manuscript has been carefully reviewed by an experienced editor whose first language is English and who specializes in editing papers written by scientists whose native language is not English.

Results

2. Consider rounding your correlation numbers to the hundredths place – I’m not sure the utility of the r-values to the thousandths place – 0.796 and 0.80 mean the same thing to the reader.
   [Response] Thanks for the suggestions. To present the detail information, we referred the paper published by the BMJ Open journal and give the thousandths place of r- and *p*- values in the paper.

3. Not sure if data from Taiwan can be generalized to all non-English-speaking countries. The authors repeatedly state that their findings are for non-English-speaking countries, and I think that statement is too broad.
   [Response] Actually, the local evidence was our intrinsic limitation of the present study. However, we present the finding of the non-English web query in Google Trends may be useful to assist the epidemic surveillance and disease control in Taiwan. We think that our initial work has the potential to develop more studies of Big Data utility in non-English-speaking countries, especially in medical sources lacking areas.

Abstract

4. Line 50 – “assess” not “access”
   [Response] We have revised the error typo.

Results

5. Page 7, line 14 – I think you mean “no forward”, not “on forward”, and “no forward” sounds weird, consider rephrasing.
   [Response] We have revised the error typo and used “lag” to replace “forward” in the paper.

Discussion

6. The authors need to discuss why some flu-related symptoms (fever, cough) are highly correlated with flu-like illnesses, but not other symptoms (runny nose)? It seems like the lay-person would be searching all of these.
   [Response] Thanks for the suggestion. We think that search behaviors are directly relative to what they concern about. During epidemic outbreaks, non-professional people use the queries of the obvious symptoms derived Flu to search the information on the web. Thus, based on our evidence, certain queries showed a higher correlation with epidemic data (e.g., common cold, fever, and cough in ILI), which may reflect what people concerned about and their web search behaviors in the epidemic outbreak.

7. The authors make this statement: “The web user’s education level, economic situation, cultural and language backgrounds can influence the local habits of Internet searchers.” But also need to clarify that Google Trends does not allow for any capture of demographic information of the Google users. This is a big limitation of google trends.
   [Response] This is an appropriate comment. We also noted Google Flu Trends has been failed to provide accurate predictions concerning influenza-like-illness (ILI) cases. Some possible intrinsic limitations regarding the use of big data on epidemic disease surveillance should be concerned in the study. However, our rationale for the present study is to combine the web query data and epidemic conditions to develop a useful real-time surveillance tool. The initial evidence does support our hypothesis and encourage us to do further study in “infodemiology”. To well explain the limitation of the study, we added related descriptions in the discussion section.

8. Page 11, Like 28 – In the US, it is “Twitter”, not “Tweets” for the social media site.
   [Response] We have revised the social media term.

Figures 1-4

9. These could easily be combined into a 4-panel single figure.
   [Response] Thanks for the suggestion. We think the separated figures may help the reader to read our results; however, we respect the editor’s comment for the final publication.

10. Seeing the temporal association is nice, but it’d be useful to use other health-related search terms that are not related to influenza (like myocardial infarction or diabetes, that lack seasonality) to show discriminant ability of Google Trends. Perhaps all health-related search terms increase around week 15-17 of your study period.
    [Response] We sincerely appreciate your comments. Based on the suggestion, we have further analyzed the correlation of non-seasonality queries with the ILI-data. As can be seen in the below table, these Chinese queries [心肌梗塞 (myocardial infarction), 糖尿病 (diabetes)] showed poor to moderate correlation with ILI-data. Therefore, these showed that the query terms we used in the present have a good discriminant ability for the analysis.

Supplementary Table: Pearson correlation coefficient values for the intensity of non-influenza-related query terms in Taiwan. (from October 4, 2015 to April 2, 2016.)

心肌梗塞 (myocardial infarction)	Correlation coefficients (r-value)	.359	.357	.395	.360
	P-value (2-tailed)	.071	.073	.046	.071
糖尿病 (diabetes)	Correlation coefficients (r-value)	-.408	-.344	-.300	-.358
	P-value (2-tailed)	.039	.085	.136	.073

Reviewer: 3

Reviewer Name

Jianhua Liu

Institution and Country

the Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China

Please state any competing interests or state ‘None declared’:

None

Please leave your comments for the authors below

There could be a lot of bias for the model in this study. Personally, I do not agree that by analyzing only the google trend could represent the epidemic outbreaks. For instance, if there is a hot topic recently about the malaria in Africa, the trend of searching related topic may increase in Taiwan as well. But we cannot conclude the prevalence of disease increase in Taiwan base on this information. Moreover, those confounders like education level and age were not included as confounders in this study, which should have both influence for outcome and expose. Therefore, I cannot interpret the result that google trend represent the epidemic disease base on this study.

[Response] We sincerely appreciate your comment on our paper. There are still some possible intrinsic limitations regarding the utility of big data on epidemic disease surveillance. Thus, in order to develop the web-based epidemic surveillance system, algorithms and computational techniques, which are built and rely on the analysis, still need to be carefully refined, tuned, and calibrated to avoid the overfitting risk in big data inference. As you mention above, we agree that the query search intensity may not reflect the real prevalence of the disease in the local area, due to the web opinion or international news. However, we present the finding of the non-English web query in Google Trends may serve as a useful tool to assist the epidemic surveillance and disease control in Taiwan. We think that our initial work has the potential to develop more studies of Big Data utility in non-English-speaking countries, especially in medical sources lacking areas.

Reviewer 4

Samuli Pesälä

Institution and Country

University of Helsinki, Helsinki, Finland

General comment

This study describes the non-English searches from Google Trends to assess epidemic diseases and public opinion through popular search behavior. The study compares epidemic data on ILI and EN71 to query data from Google Trends. The results suggest that Google Trends serve as a good surveillance tool for epidemic outbreaks in non-English countries. The study has public health impact in terms of applying traditional register-based data and Google Trends on infectious diseases (ILI, EN71) in order to enhance disease surveillance in the era of Internet. Yet the manuscript needs major revisions.

[Response] Thanks for the kindly suggestions and comments for our work.

Major comments

1. Introduction: Please include more information on enterovirus infection, its seasonality and epidemics in the Introduction section. Also, a short introduction is needed explaining the seasonality of influenza.
   [Response] Thanks for the kindly suggestion. We have additionally described the seasonal features of these two infectious diseases in the introduction section.

2. Page 4, lines 11-13: Was there a reason why you collected data on influenza in 10/2015-4/2016, but enterovirus in 1-12/2012? Data availability? Please shortly mention the reason.
   [Response] Thanks for the comments. In this studies, we focused on using google trends to monitor the epidemic disease and the public opinion. During the period of the weeks (from October 4, 2015, to April 2, 2016), an influenza outbreak, a natural disaster, and an earthquake, were occurred in Taiwan. Thus, the period was suitable to be the target, which we estimated whether the Google trends was suitable to be the surveillance of the epidemic disease and the public opinion. Furthermore, we have already described the limitation of our study in the discussion section. In the future, we will persist to develop more prediction models based on the long term Internet big data to improve the effectiveness of epidemics surveillance in Taiwan.

3. Methods, Page 4, lines 29-31: I suppose that enterovirus analysis included a laboratory test which tested enterovirus infection among patients, or was enterovirus infection based on clinical symptoms? Please clarify.
   [Response] Clinically, a well-trained physician can often diagnose patients infected with enterovirus based on the symptoms. Although some laboratory tests were established to confirm acute enterovirus infection, such as cell culture, serology, and PCR, those remained to exist the time consumed limitation.
   However, here we presented the “enterovirus analysis”, which indicated the epidemiological surveillance data obtained from Taiwan CDC and we conducted to evaluate the correlation of those open data with the Google Trends search relative intensity. To avoid confusion, we made the minor revision of the description, using “For the enterovirus survey” instead of “For the enterovirus analysis”.

4. Results, Page 5, lines 37-38: “The total of 8 queries related to influenza (Table 1)”. There are 10 query terms in the table. Which terms related?

[Response] Thanks for kind reminding. Based on our approach, there are 10 queries related to ILI, dividing into 4 categories in the present study. I have corrected the total number of related queries in the text.

5. Page 7, lines 14-15: “Thus, the group C keywords research relative intensities”. What and where is group C? This is not mentioned anywhere else in the manuscript.
   [Response] This is our proofreading error. We have removed the term “Group C” in the result section.

6. Figures 1-4, Pages 16-19: Weeks are reported “relatively” (weeks 1-25, i.e. 6 months), not real weeks. Now it seems that an influenza epidemic occurs during weeks 15-23 (April-June). According to other research, influenza weeks in Taiwan should occur something like between weeks 50-14. These should be revised into real weeks. However, in Figure 5, during the year 2012 (weeks 1-52), the peak week (wk 21-26) of an enterovirus epidemic seems to be in the summertime. This seasonality should be explained in the Introduction.
   [Response_6.1] We have revised the title of the X-axis to present the real weeks in Figure 1-4 and Figure 6. Also, we simultaneously revised the description of the week in the text.
   [Response_6.2] The seasonality of the enterovirus infection has been additionally described in the introduction section.

7. Figure 4, Page 19: There is “death of pneumonia and ILI patients” visualized with red vertical bars. This is a little bit unclear. Please choose plural (“deaths”) as shown in Table 2 (“weekly deaths from pneumonia and ILI”). Had these patients had ILI diagnoses before they died of pneumonia?
   [Response] Thank you for your suggestion. We have revised to increase the contrast of the vertical bar's color to clearly represent the weekly deaths of pneumonia and ILI patients in Figure 4. Furthermore, according to the criteria of the real-time monitoring system established by the Taiwan CDC, the definition of “weekly deaths of pneumonia and ILI patients” is the number of patients who died for both pneumonia and influenza-like illness syndromes.

8. In public opinion estimation, why did you choose the general public’s query term “ECMO” to be compared with positive influenza tests? Please explain.
   [Response] ECMO is the well-known emergency medical equipment for most Taiwanese. Because of the new reports of severe influenza cases by mainstream media during the outbreak period, the search term “ECMO” increased the related intensity of Google trends. Thus, we used the certain terms, such as influenza, ECMO, and Tamiflu, to estimate whether the query data reflect the public opinion during epidemic outbreaks.

9. Page 22: Please omit the results from your future studies (Figure in Multimedia Appendix 2, Portuguese keywords).
   [Response] Thanks for your suggestion. We have removed this section in the text.

Minor comments

10. In the Abstract (strengths and limitations), “EN71” should be written out when mentioned for the first time in the text, or replaced with “enterovirus infection”.
    [Response] We have revised the description in this section.

11. In the Abstract, strengths and limitations: “Apart from Google trends, we need to combine more

social media to comprehensively analysis the epidemic information through web-related behaviors". This sentence should be clarified.

[Response] To more explain this sentence, we revised the description in the section. Based on our approach, we actually need more big data, such as social media, local meteorology, and resident consumptive behaviors, to comprehensively analysis the epidemiologic information and predict the outbreak in time.

12. Results: Page 6, lines 33-34: There is background information on enterovirus ("Enterovirus 71 (EN71) was first identified in California, USA, in 1969."). I think this should be included in the Introduction section, not Results section.

[Response] Thanks for the comment. For the convenience of the reader, we tend to keep the background information at the beginning of this paragraph, to serve as the rationale of the analysis.

13. Figure 1 title, Page 16: "…keywords research relative intensity…" Is this a defined/formal term or do you mean "keywords search relative intensity"? In Figure 1, there is "Google Trends search relative intensity" on the right hand side.

[Response] To consistently reveal the information in the figure and figure legend, we revised all the figure legends to present the same content as the title of the right Y-axis in figures.

14. Page 7, lines 17-18: "…was 4-weeks log between the Internet query data and epidemic advance." Do you mean that there was a 4-week delay in queries? Then, I think it is a "lag".

[Response] Yes, here we indicated the delay period between the Internet query data and the epidemic advance. Thanks for the suggestion, we have revised the error typo.

15. Table 3: Page 9, lines 10-11: A typo "entervirus", should be "enterovirus"

[Response] We have revised the error typo.

16. Figure 5, Page 20: Two terms are used "enteroviruses infection" and "enterovirus infection". Please choose "enterovirus infection".

[Response] We have revised this term in Figure 5 and the figure legend.

17. Supplementary material, Page 12: In the table, a typo "ECOM", should be "ECMO".

[Response] We have revised the error typo.

18. Please check the initial letters regarding "Figure" and "Table" and make sure that these are consistently uppercase throughout the text if you refer to the Figures and Tables where you show the data. Also, please check that "Internet" is uppercased in the text.

[Response] We have checked the initial letters of Figure, Table, and Internet in the text and revised them in uppercase.

VERSION 2 – REVIEW

REVIEWER	Aaron Secrest University of Utah, United States

REVIEW RETURNED	03-Apr-2020

GENERAL COMMENTS	I feel the authors have adequately addressed all of my concerns and the English writing is much improved.

REVIEWER	Samuli Pesälä University of Helsinki, Finland
REVIEW RETURNED	07-Mar-2020

GENERAL COMMENTS	The authors have answered my questions properly and revised the text as suggested. I am happy with the revisions.