REVIEW

The most cited articles on the topic of health behaviors in Google Trends research: a systematic review

Wei Chih Kan¹,² Tsair Wei Chien² Hsien Yi Wang¹,⁴ Willy Chou⁵,⁶∗

Abstract: Background: Over the past decade, the use of Web-based data in public health issues has been proven useful in assessing various aspects of human behavior. Google Trends is the most popular tool to gather such information and has been applied to several topics with the most focused subject related to health and medicine. However, the most cited articles and the popular medical subject headings (MESH terms) on health behaviors in Google Trends research remain unknown. The web-based behavior requires to monitor and analyze on-line data for examining actual human behavior to predict and even prevent health-related issues that constantly arise in daily life. Objective: This systematic review aimed at reporting and further presenting the most cited articles and the popular MESH terms on health behaviors in Google Trends (infodemiology) researches in health-related topics since 2009 to provide an overview of the topic burst for future research on the subject of health behavior. Methods: Following the Meta-Analyses guidelines for selecting studies, we searched for the term “Google Trends [Title]” in PubMed databases since 2009, applying specific criteria for types of journal articles. A total of 86 published papers were extracted, excluding those that did not fall inside the topics of health and medicine or the selected article types. We then further categorized the published papers according to MESH terms using social network analysis (SNA) and selected the most cited articles that related to the health behavior in Google Trends. Results: The most cited articles are those from the US in 2009 (PMID= 19845471 cited 88 times) and the UK in 2013 (PMID= 23619126 cited 74 times). The MESH term represented by Internet earns the highest impact factor (IF) and presents significantly different among term clusters (F (3,20)=15.79, p<0.001). The most number of citing journals is from PloS One. The most number of author affiliations is from the US. Conclusion: The monitoring of online queries can provide insight into human behavior, as the phenomenon is significantly and continuously growing at present and in the future for assessing behavioral changes in health topics.

Keywords: health behavior, infodemiology, Google Trends, medical subject headings, social network analysis

1 Introduction

Big data are characterized by the eight Vs[¹–⁵](i.e., volume, variety, velocity, veracity, value, variability, volatility, and validity) and have shown great potential in forecasting and better decision making. Handling these data with conventional ways is inadequate[⁶] and hence requires novel approaches and methods applied to healthcare research.[⁷] The online search queries have been popular in big data analytics for academic research.[⁸,⁹] The use of search traffic data from web-based sources can assist in facilitating a better understanding the Web-based behavior and behavioral changes.[¹⁰] Online search traffic data have been deemed as a good analyzer of internet behavior, while Google Trends[¹¹] acts as a reliable tool in predicting changes in human behavior and as an accurate measure which is the publics interest.[¹²] What are the most popular topics in health behavior using Google Trends remain unknown. The monitoring of web-based activity (or health behavior) can be thus a valid indicator of public behavior.[¹³]


Copyright: © 2018 Willy Chou, et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Healthcare is one of the top fields to which big data are widely applied\cite{14,15} with many publications showing a high increase.\cite{16} Researchers have also placed their efforts on examining Web-based search queries for health behavior and medicine related topics.\cite{17} However, few were reported on the most cited articles and topics which were searched in Google Trends. Which author affiliations or regions searching health topics in Google Trends research are most prominent in the world are required to investigate.

As the use of Google Trends in examining human behavior is relatively novel, the search health topics of behaviors are constantly arising. Many topics have been examined, such as epilepsy\cite{18,19} cancer,\cite{20} thrombosis,\cite{21} silicosis,\cite{22} and various medical procedures including cancer screening examinations,\cite{23,24} bariatric surgery,\cite{25} and laser eye surgery.\cite{26} We are thus interested in exploring (1) Which topics are most searched in healthcare; (2) which types of MESH terms are most characterized in Google Trends; (3) which articles are cited most in the past; (4) which regions in research affiliations are most in use of Google Trends in the world.

2 Methods

2.1 Data Collection

This review aimed to include all articles on the topics of health and medicine that have used Google Trends in the literature. We searched for the term “Google Trends” in article Title on PubMed\cite{27} since 2009, and following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines, the total number of publications included in this review was 86.

2.2 Search in Pubmed

First, we searched for the keyword “Google Trends” in the “Abstract-Title-Keywords” field for the journal articles. The first two articles using Google Trends were begun in 2009. The search returned 96 publications.

2.3 Social network analysis and Pajek software

Social network analysis (SNA)\cite{28} was applied to explore the pattern of entities in a system using the software of Pajek.\cite{29} In keeping with the Pajek guidelines, we defined an author (or paper keyword) as a node that is connected to other nodes through the edge (or say the relation). Usually, the weight between two nodes is defined by the number of connections.

Centrality is a vital index to analyze the network. Any individual or keyword lies in the center of the social network will determine its influence on the network and its speed to gain information.\cite{30,31}

2.4 Tasks to report the feature on health behaviors in Google Trends research

2.4.1 What are the most number of author-defined keywords in the network

Author-defined keywords were collected and analyzed using SNA to separate clusters. The most number of betweenness centrality in each cluster is presentative which is the most occurred frequency in the respective cluster.

2.4.2 What are the most number of MESH terms in the network

Similarly, the MESH terms were used to characterize the feature for each cluster, which is significantly different from the author-defined keywords and actually and objectively represents the feature clusters of articles using Google Trends. Furthermore, all articles were grouped into the respective MESH terms through the maximum likelihood method (MLM). That is, the highest score among MESH cluster is the one article belonged to. Whether the types of MESH clusters are different in the number of the citation will be examined using the one-way ANOVA. The article citations were retrieved from the Pubmed Central (PMC). The most cited articles will be reported in this study.

2.4.3 What are the most number of citing journals related to the searching topics in Google Trends research

After retrieving the citing journals for articles from PMC, the association of journal citation can be analyzed using SNA. The most number of betweenness centrality were given to the journal representing the designated cluster of citing journals.

2.4.4 What are the most influential research regions regarding health behavior using Google Trends

The most influential research regions based on the order of author affiliations using the authorship-weighted scheme.\cite{32} That is, the first (i.e., primary) and the last (i.e., corresponding) authors gain higher weighted contributions to the given article.\cite{33,34} The visual display will be shown with a dashboard on Google Maps.

2.5 Statistical tools and methods used in this study

The Kendalls coefficient of concordance (W)\cite{35} was computed to examine the internal consistency (IC) of the data (i.e., the four indices of h, g, x, and L indexes as well as author impact factor (AIF)) related to MeSH clusters. If the agreement is accepted by the statistical al-
phas level (<0.05),\textsuperscript{[42]} the following analysis of one-way ANOVA for inspecting the difference in mean of indices is meaningful.

3 Results

3.1 Task 1: the most number of author-defined keywords in the network

Author-defined keywords displaying the top three representatives of clusters are Google trends, digital epidemiology, and infodemiology (Figure 1). It implies that many are methodology except for the only one topic of cosmetic surgery in the network.

Figure 1. Author-defined keywords dispersed on a dashboard

3.2 Task 2: the most number of MESH terms in the network

As to the MESH terms, we see the MESH terms constituting the top five topics of the internet, trends, statistics & numerical data, information seeking behavior, and web browser (Figure 2), which indicate that all articles regarding health behaviors in Google Trends can be formed by the five types of clusters.

The relevant bibliometric indices (i.e., the four of h, g, x, and L indexes as well as AIF) to the MESH clusters are closely associated with each other (i.e., having high correlation coefficients). The Kendalls W is 0.67 with three degrees of freedom (χ\textsuperscript{2}=12.05, p=0.01), indicating data with acceptable internal consistency (IC). The MESH term represented by Internet earns the highest impact factor (IF) and presents significantly different among term clusters ( F(3,20)=15.79, p<0.001).

Figure 2. MESH terms dispersed on a dashboard

3.3 Task 3: the most number of citing journals

The most number of citing journals is from PloS One (Figure 3) followed by the journals of JMIR Public Health Surveill and J Med Internet Res. The most cited articles are those from the US in 2009 (PMID= 19845471 cited 88 times) and the UK in 2013 (PMID= 23619126 cited 74 times).\textsuperscript{[43]}

Figure 3. Clusters of citing journals related to the topic of health behaviors on Google trends

3.4 Task 4: the most influential research regions

The most number of author affiliations is from the US (Figure 4) followed by Italy and the UK as well as China, indicating the most influential research regions using Google Trends on healthcare are from the US, Europe, and China.

Figure 4. Clusters of author affiliations related to the topic of health behaviors on Google trends

4 Discussion

This study found that (1) the most cited articles are those from the US in 2009 (PMID= 19845471 cited 88 times) and the UK in 2013 (PMID= 23619126 cited...
74 times); (2) the MESH term represented by Internet earns the highest impact factor (IF) and presents significantly different among MESH clusters (F (3,20)=15.79, p<0.001); (3) the most number of citing journals is from PloS One; (4) the most number of author affiliations is from the US.

Referring the previous study\[13\] which assessed the methods, tools, and statistical approaches in Google Trends Research, a total of 23.1% (24/104) studies used Google Trends data for examining seasonality, while 39.4% (41/104) and 32.7% (34/104) used correlations and modeling, respectively. A few around 8.7% (9/104) used for predictions and forecasting in health-related topics. All the 104 examined papers but two included data visualization to present the study results. For instance, a worldwide map examined country for assessing health and medicine related issues and found that the US data have been employed in the most (60) studies, while other countries include the UK (15), Australia (13), Canada (9), Germany (8), and Italy (7). The results are somewhat different from the visualization of our study (i.e., the US, Italy, the UK and China has shown in Figure 4. It is because we emphasize the author collaboration instead of the regions applied by Google Trends.

Many studies have employed Google Trends for visualizing the changes of interest or discussing peaks and spikes\[34–47\] such as the search volumes for related terms, terms related to the studied topic\[48,49\] and the related internet searches\[50\]. Others include the reporting of the polynomial trend lines and the investigating of statistically significant differences in yearly increases\[51\]. Even Google Correlate has also been used to explore related terms\[52\] we have not found any applying author-defined keywords or MESH terms to visualize the related terms, as we show them in Figure 1 and Figure 2.

The vast majority of studies using Google Trends in health assessment have employed data visualization, such as figures, maps, or screenshots. The most popular way is correlating them with official data on disease occurrence, spreading, and outbreaks. For instance, the assessment of suicide tendencies and (prescription or illegal) drug-related queries has been popular over the past years. However, the gap in the existing literature is the lack in use of Google Trends for predictions and forecasting in health-related topics which can benefit the general public by using and analyzing web-based data to provide insight to better assess health issues and topics in healthcare settings.

Finally, we particularly provided citation analysis to illustrated the most number of citing journals and proved that the MESH term of Internet earns the highest impact factor (IF) and presents significantly different among MESH clusters (F (3,20)=15.79, p<0.001). Another feature in this study is about the usage of Kendall’s coefficient of concordance (W) which told us the bibliometric indices closely related to each other in Table 1, which is also rarely seen in the literature, particularly for data not following the normal distribution.

### 5 Limitations and Future study

The interpretation and generalization of the conclusions should be cautious. First, the data were extracted from Medline. It is worth noting that any generalization should be made in similar fields of paper related contents.

Second, several studies have used other sources of big data, namely, Google News \[53–55,58\] Twitter \[56,57,59\] Yandex, \[60\] Baidu, \[61\] Wikipedia, Facebook and Google+, and YouTube. Even Google is the most popular search engine, other Web-based sources are used or even preferred to Google in some regions (e.g., in China) which is worthy of study in the future.

Third, although the data were carefully extracted from Medline and were professionally dealt with in every link- age, the originally downloaded contexts with some errors in symbols might affect the resulting reports.

---

**Table 1. Features of MESH clusters across bibliometric indices**

<table>
<thead>
<tr>
<th>MESH cluster</th>
<th>Output</th>
<th>Cited number</th>
<th>IF</th>
<th>h</th>
<th>g</th>
<th>x</th>
<th>L</th>
<th>Ag</th>
</tr>
</thead>
<tbody>
<tr>
<td>internet</td>
<td>20</td>
<td>374</td>
<td>18.7</td>
<td>9</td>
<td>13</td>
<td>15.44</td>
<td>19.34</td>
<td>28.62</td>
</tr>
<tr>
<td>trends</td>
<td>11</td>
<td>13</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5.74</td>
<td>6.2</td>
<td></td>
</tr>
<tr>
<td>statistics &amp; numerical data</td>
<td>16</td>
<td>55</td>
<td>3.44</td>
<td>4</td>
<td>7</td>
<td>4.69</td>
<td>7.42</td>
<td>7.29</td>
</tr>
<tr>
<td>information seeking behavior</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>1.41</td>
<td>1.41</td>
<td>7.29</td>
</tr>
<tr>
<td>web browser</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Mean**

<table>
<thead>
<tr>
<th>IF</th>
<th>h</th>
<th>g</th>
<th>x</th>
<th>L</th>
<th>Ag</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>92.8</td>
<td>5.25</td>
<td>5.5</td>
<td>8</td>
<td>6.31</td>
</tr>
</tbody>
</table>

**Correlation Coefficient**

<table>
<thead>
<tr>
<th>IF</th>
<th>h</th>
<th>g</th>
<th>x</th>
<th>L</th>
<th>Ag</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.98</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.04</td>
<td>0.89</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.08</td>
<td>0.07</td>
<td>0.07</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.08</td>
<td>0.94</td>
<td>0.09</td>
<td>0.09</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0.99</td>
<td>0.97</td>
<td>0.97</td>
<td>0.95</td>
<td>0.94</td>
<td>1</td>
</tr>
</tbody>
</table>

Kendall’s consistency:

- W = 0.67
- Z = 12.05
- df = 3
- prob. = 0.01

**Cluster ANOVA:** F(3,20)=15.79, p<0.001
Fourth, there are many algorithms used for SNA. We just applied community cluster and betweenness centrality with weighted degrees in Figures. Any changes made with the specific algorithm will present different pattern and inference making.

Fifth, the social network analysis is not subject to the Pajek software we used in this study. Others such as Ucinet [62] and Gephi [63] are suggested to readers for use in the future.

6 Conclusion

This review consists of the studies published from 2009 to 2018 on Google Trends research in the PubMed databases based on the selected criteria. This review aimed to provide a point of reference for future research in health behaviors using Google Trends. Google Trends data are commonly used in infodemiology research and have been shown to empirically correlate with official health data in many topics. Using Google Trends become increasingly popular in health behavior assessment that can be crucial in monitoring and analyzing seasonal diseases as well as epidemics and outbreaks in healthcare settings.

7 List of abbreviations

AIF: author impact factor  
BC: Betweenness centrality  
IC: internal consistency  
IF: impact factor  
MESH: medical subject headings  
PMC: Pubmed Central  
SNA: Social network analysis

8 Competing interests

The authors declare that they have no competing interests.

9 Authors contributions

WC conceived and designed the study, TW performed the statistical analyses and were in charge of dealing with data. HY and TW helped design the study, collected information and interpreted data. WC monitored the research. All authors read and approved the final article.

References

https://doi.org/10.1126/science.1200970  
https://doi.org/10.1016/j.ins.2014.01.015  
https://doi.org/10.1016/j.bdr.2015.01.006  
https://doi.org/10.1016/j.ijpe.2014.12.031  
https://doi.org/10.1016/j.dss.2013.08.008  
https://doi.org/10.1016/j.ijinformat.2014.10.007  
https://doi.org/10.25082/AHB.2018.01.005  
https://doi.org/10.25082/AHB.2018.01.004  
https://doi.org/10.1016/j.techfore.2013.04.013  
https://doi.org/10.1093/ijpor/edq048  
https://doi.org/10.2196/jmir.9366  
https://doi.org/10.1093/nsr/nwt032  
https://doi.org/10.5213/inj.2014.18.2.50


[38] Belikov AV and Belikov VV. A citation-based, author- and age-normalized, logarithmic index for evaluation of individual researchers independently of publication counts. F1000Research, 2015, 4: 884. https://doi.org/10.12688/f1000research.7070.1


[40] Pan RK and Fortunato S. Author Impact Factor: tracking the dynamics of individual scientific impact. Scientific Reports, 2014, 4: 4880. https://doi.org/10.1038/srep04880


