CAN INTERNET SEARCHES PREDICT AN OUTBREAK?
A Literature Review on Nowcasting with Google Trends

Daniel Bil

Abstract
Aim: The aim of this literature review is to provide an introductory overview of the use of Google Trends for public health surveillance, its strengths and limitations, and areas for further research.

Methods: A literature search was conducted through PubMed to identify publications which used Google Trends data (GTD) or an existing GTD model as a source to investigate an infectious disease.

Results: Google Trends has been used to model the incidence of a range of diseases, including influenza, dengue fever, HIV, pertussis, and malaria, with varying degrees of accuracy. Models frequently correlate with reported case data at a negative time lag, providing warnings of outbreaks earlier than traditional systems. Case data can be incorporated into models for greater accuracy. News/media bias, a small population size, and various sociodemographic factors are recurring themes noted to reduce accuracy.

Discussion: With the potential to monitor disease incidence in real time and improve existing modelling solutions, Google Trends represents an exciting new frontier for epidemiology and public health. However, these tools should be positioned as an adjunct to traditional public health surveillance rather than a replacement. More real-world testing in diverse cultural settings is necessary to better understand its strengths and limitations across the digital divide.
Introduction

Since the turn of the century, the Internet has become an increasingly popular way for people to access health information. Search engines, particularly Google, have become a popular choice for many wanting to explore their symptoms and diagnoses due to its accessibility, familiarity, and ease of use. In the past few years, discussion and novel research have arisen on whether data from Internet search queries that are potentially related to a disease could be correlated with real cases. If so, these correlations could be developed into accurate predictive models, unlocking a suite of public health surveillance tools for nowcasting the incidence of a disease in real time. This emerging field has been termed ‘digital epidemiology’. [1]

Google Trends

Google Trends (GT) is a freely available online tool developed by Google LLC which allows users to input search queries and retrieve data on their relative search volume (RSV) through the Google search engine over a given period. Searches can be filtered by location and visualised on a line chart and heat map. RSV is reported as a number from 0-100, with 0 representing no searches and 100 representing the peak search volume within the timeframe. Data can then be exported and analysed further with external software. [2]

In 2009, Ginsberg et al [3] used search query data from Google Trends to develop a real-time influenza surveillance tool, representing the first major contribution in the field. The model used a selection of 45 weighted search term RSVs to estimate the number of US physician visits for influenza-like illness (ILI) in a given week, as a proportion of all physician visits. The model was fit using historical data from 2003 to 2007 and tested against out-of-sample data from the 2007-08 US influenza season with highly correlated (ρ=0.97) results.

The subsequent launch of Google Flu Trends (GFT) and Google Dengue Trends, built from a similar methodology, [4] catalysed a surge of academic interest in the new field. However, in 2009 and 2013, GFT suffered setbacks related to erroneous predictions and support for the service ended in 2015. [5] Despite this, its impact on digital epidemiology remains relevant, and the academic interest it has sparked continues to persist.

Aim

The aim of this literature review is to provide an introductory overview of the use of Google Trends for public health surveillance, its strengths and limitations, and areas for further research.

Method

A literature search was conducted in PubMed for the terms “internet”, “search”, and “trend(s) OR query/queries” in the title and/or abstract. Relevant articles were identified based on their title and abstract, as well as appraised based on their method and findings. Articles were excluded if they did not use Google Trends data (GTD) or an existing GTD model (such as Google Flu Trends or Google Dengue Trends) as a source to investigate an infectious disease.

Results

The literature can be divided into two broad categories: correlation studies and modelling studies. Correlation studies can identify relationships between GTD and case data, but do not allow for real-time estimates or predictions to be made. Table 1 lists a selection of significant correlation studies providing contributions to the field.

<table>
<thead>
<tr>
<th>Year of publication</th>
<th>Author</th>
<th>Disease</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>Cho et al [6]</td>
<td>Influenza</td>
<td>South Korea</td>
</tr>
<tr>
<td>2015</td>
<td>Domnich et al [8]</td>
<td>Influenza</td>
<td>Italy</td>
</tr>
<tr>
<td>2018</td>
<td>Shah et al [9]</td>
<td>Rotavirus infection</td>
<td>United States, United Kingdom, Mexico</td>
</tr>
<tr>
<td>2018</td>
<td>Verma et al [10]</td>
<td>Malaria, dengue fever, chikungunya, enteric fever</td>
<td>India</td>
</tr>
</tbody>
</table>

Table 1: GTD correlation studies.
In contrast, modelling studies attempt to describe the relationship between GTD and disease incidence, allowing for real-time nowcasting using out-of-sample data. Table 2 lists a selection of nowcasting studies identified in this review.

Another subset of studies explore applications of existing GTD models in new settings, namely Google Flu Trends [24-27] and Google Dengue Trends. [28-30]

The main measures of correlation strength used in the literature were Pearson’s correlation coefficient (r), Spearman’s rank correlation coefficient (ρ), and coefficient of determination (R²). For ease of comparison, R² values have been converted to r by taking its square root. Studies typically analysed data at a daily or weekly level, reflecting the temporal resolution provided by Google Trends.

**Strengths**

GTD models frequently correlate with reported case data with a negative time lag, indicating changes in search volumes may take place earlier than corresponding changes in reported cases. Ginsberg et al [3] hypothesised this could be due to the time required by traditional surveillance systems to compile notification data, delaying their publication.

This suggests GTD models could provide an n-period ahead forecast for changing incidence patterns relative to traditional systems, providing warnings of outbreaks days to weeks in advance. Samaras et al [13] showed how their model could forecast, 5 weeks in advance, the peak of the 2012 scarlet fever resurgence in the UK. Verma et al [10] showed that the strongest correlations between GTD and chikungunya, dengue fever, malaria, and typhoid fever in India were present at a -2 to -3 week lag. McGough et al [18] showed how GTD models could more accurately predict the spread of the Zika virus than case data models 2-3 weeks into the future. Santangelo et al [11] showed the correlation between measles-related searches and infections in Italy was strongest at a lag of -3 weeks (ρ>0.80).

In addition, multiple studies have found that nowcasting models built from both GTD and historic data may attain greater accuracy. Pries and Moat [15] found that nowcasts of influenza levels based on GTD and case data had a mean absolute error 16.0%-52.7% lower than models based on case data alone. Gluskin et al [28] found that accounting for climate factors (maximum temperature and precipitation) alongside Google Dengue Trends data improved accuracy compared to search data alone (r=0.90 and r=0.82, respectively). However, McGough et al [18] identified the possibility that GTD alone may outperform GTD coupled with case data if the case data is flawed.

**Limitations**

News and media bias are a recurring theme throughout the literature. A celebrity diagnosis, drug recall, public health campaign, or other event resulting in unusually high media coverage can increase the volume of search queries from those who are not ill. Google Flu Trends famously overestimated the peak

<table>
<thead>
<tr>
<th>Year of publication</th>
<th>Author</th>
<th>Disease</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>Ginsberg et al</td>
<td>Influenza</td>
<td>United States</td>
</tr>
<tr>
<td>2012</td>
<td>Samaras et al</td>
<td>Scarlet fever</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>2013</td>
<td>Ocampo et al</td>
<td>Malaria</td>
<td>Thailand</td>
</tr>
<tr>
<td>2014</td>
<td>Preis &amp; Moat</td>
<td>Influenza</td>
<td>United States</td>
</tr>
<tr>
<td>2017</td>
<td>Zhang et al</td>
<td>Pertussis</td>
<td>Australia</td>
</tr>
<tr>
<td>2017</td>
<td>Kandula et al</td>
<td>Influenza</td>
<td>United States</td>
</tr>
<tr>
<td>2017</td>
<td>McGough et al</td>
<td>Zika</td>
<td>Latin America</td>
</tr>
<tr>
<td>2017</td>
<td>Samaras et al</td>
<td>Influenza</td>
<td>Greece and Italy</td>
</tr>
<tr>
<td>2018</td>
<td>Young et al</td>
<td>Syphilis</td>
<td>United States</td>
</tr>
<tr>
<td>2018</td>
<td>Young et al</td>
<td>HIV</td>
<td>United States</td>
</tr>
<tr>
<td>2018</td>
<td>Oren et al</td>
<td>RSV infection</td>
<td>United States</td>
</tr>
<tr>
<td>2019</td>
<td>Arehart et al</td>
<td>Pertussis</td>
<td>United States</td>
</tr>
</tbody>
</table>

Table 2: GTD modelling studies.
of the 2013 US flu season by almost double that reported by the CDC, which was believed to have been caused by unusually widespread media coverage. [5] This is mirrored by experiences during the 2014 Ebola epidemic, where GTD in the three countries showed moderate (r=0.640 in Liberia, p<0.001) to insignificant (r=0.232 in Guinea, p=0.07) correlation with case data, with search volumes disproportionately influenced by news coverage locally and overseas.[7] Ho et al [29] noted an increase in the volume of certain dengue-related searches when Michael V, a well-known Filipino celebrity, was diagnosed with the disease in 2013.

The degree to which regional incidence can be resolved is also a limitation to the usefulness of GT data. Gluskin et al [28] found that, although Google Dengue Trends showed strong correlation with official data in Mexico at the national level, its reliability varied enormously from state to state (r=0.01-0.94), with states with low incidence performing worse. Similarly, Strauss et al [30] found the tool performed well as an average over the entire period studied (r=0.87), but its performance suffered during non-epidemic weeks (r=0.65) and the low-incidence seasonal trough in March (r=0.48). Zhang et al [16] also noted significant variation at the state/territory level in Australia (p=0.17-0.76) when analysing correlations in pertussis incidence. Arehart et al [23] found while their US pertussis model correlated significantly with case data at a national level, their state models could not track case data in some states. Ginsberg et al [3] found that their national-level influenza model was less accurate, though still strongly correlated, when applied to an individual state, Utah (p>0.90).

Sociodemographic factors such as age, birth rate, and education attainment have been noted to generally impact correlation strength, however these findings appear to be less consistent.[16,23,28] Conversely, Gluskin et al [28] found socioeconomic factors including Internet access did not strongly impact the accuracy of Google Dengue Trends.

Discussion
The potential for Google Trends and other online sources to identify and predict infectious disease outbreaks presents an exciting new frontier for epidemiology. The list of use-cases is vast, from providing earlier responses to outbreaks, to adequate staffing of clinics and emergency departments, and improving the quality of official surveillance where infrastructure is lacking or absent.

However, much of the existing data in this field is exploratory in nature. More testing against out-of-sample data and trialling GTD models in real-world scenarios is necessary to improve our understanding of digital epidemiology and its capabilities.

Areas for further investigation
It would seem GTD models are best suited to diseases with a high baseline incidence relative to their news/media coverage. Sociodemographic and population factors, such as size, internet penetration, average age, average education attainment, and information-seeking behaviours all also appear to impact their performance. This may help to explain why certain use-cases, such as influenza and HIV in the United States,[3,21] have been more promising than others, such as Ebola in West Africa or plague in Madagascar.[7,12] Further research into this aspect of digital epidemiology is warranted.

GTD models also appear to be highly context sensitive, given their variability in performance between countries, states, and even the same population over time. For the most accurate results, models will likely need to be regularly recalibrated to best fit the underlying population’s search behaviour. For this reason, further investigation into the use of GTD in culturally diverse settings, particularly in poorer and non-Western nations, is necessary to bridge the digital divide.

Issues within the literature
Given the large proportion of positive findings and the novelty of this area, the possibility of publication bias should be entertained, especially as analyses with GT are generally fast and easy to conduct. Moreover, GTD models can only be as accurate as the surveillance data on which they are trained and measured, and it is logical that any biases present in these data will be reflected in GTD models. Given this interdependency, Internet-based tools are positioned as a possible adjunct to traditional surveillance tools, rather than a replacement.

Several studies lacked thorough documentation or justification of their methodology, such as the search terms used or location(s)/time period queried, making it difficult to draw meaningful comparisons.[31] This points to a need for the adoption of a consistent methodology for conducting research
into digital epidemiology, such as those proposed by Nuti et al [31] and Mavragani & Ochoa [32]. This would be a significant step forward in improving results’ validity and replicability.

Conclusion
Google Trends is a fascinating tool which could dramatically evolve public health surveillance. The next step for this field is to move away from exploratory data and into more rigorous testing and real-world application, particularly in diverse cultural settings across the digital divide. These tools should be positioned as an adjunct to traditional surveillance rather than a replacement and should be recalibrated regularly to match the underlying population and their changing search behaviours. More research is needed into methods to correct for the impact of news/media and to adopt a consistent methodological framework for digital epidemiology.

Google and Google Trends are also not the only online sources which have potential epidemiological applications. Twitter, Baidu, and online blogs are just examples of data sources which have also been explored in academic literature and represent further potential growth for the field of digital epidemiology.

About the Author
Daniel Bil is a Doctor of Medicine student at Monash University, and Brand/IT Officer for AMSA Global Health in 2020. He is passionate about the use of technology and communications to improve public health in the digital age.

Conflicts of Interest
N/A

Correspondence
dbil0002@student.monash.edu

Acknowledgements
I wish to thank Marisse Sonido, Senior Editor at AJGH, for her invaluable guidance and support. I also thank the peer reviewer, Hyun Jae Nam, Tahlia Harper, and Cees Bil for their helpful insights.

Images
N/A

References


