Google Trends: opportunities and limitations in health and health policy research

Word count: 1,740 words (excluding abstract, references, and boxes), 2,139 including boxes

Key words: Health policy, social media, internet, health behaviours

Conflicts of interest: We have no conflicts of interest to report.

Abstract

Web search engines have become pervasive in recent years, obtaining information easily on a variety of topics, from customer services and goods to practical information. Beyond these search interests, however, there is growing interest in obtaining health advice or information online. As a result, health and health policy researchers are starting to take note of potential data sources for surveillance and research, such as Google Trends^{TM}, a publicly available repository of information on real-time user search patterns. While research using Google Trends^{TM} is growing, use of the dataset still remains limited. This paper offers an overview of the use of such data in a variety of contexts, while providing information on its strengths, limitations, and recommendations for further improvement.

Introduction

Google dominates the market for internet search engines and is so pervasive that the term “to Google” has entered everyday use in a way that none of its competitors has. In Europe it is used in 85% of internet searches[1] while in the US, it accounts for 65%.[2] In 2012, Google handled approximately 1.2 trillion searches globally, or 3.3 billion searches per day.[3] Although more recent data are difficult to obtain because of commercial confidentiality, it has been estimated that the total number rose to about 2 trillion in 2018. In an era where web searches and transactions are recorded instantaneously, this activity generates a massive volume of data whose uses are often unexpected and virtually limitless. Every search that is undertaken, and every page that is viewed, is tracked. Consequently, Google itself, along with many online content providers, such as Amazon, make extensive use of these data, tailoring advertisements and the results of searches to each user’s browsing history.

While individuals search for many things online, such as consumer goods and services, or practical information (such as opening hours or travel timetables), there is also significant search activity related to health concerns. In 2018, a report from the Pew Internet and American Life Project revealed that 80% of Internet users in the US have searched for a health-related topic online, ranging from mental health, immunizations to sexual health information.[4,5]

Health and health policy researchers are also starting to take note of the potential of these data. A PubMed search for “Google Trends” or “Google Insights” (previous version of Google Trends^{TM}) revealed an over 20-fold increase in original research articles or research letters using Google Trends^{TM} from 2009 to 2018 (**Figure 1**). While much of this information is kept secret by online providers, other elements are available to anyone. This paper offers a general guide to how Google collects and shares search engine data, how their data has been used in the past, and how health and health policy researchers can make greater use of these data in the future.

Google Trends^{TM}: What it is and how it works

Google Trends^{TM} is the principle tool used to study trends and patterns of search engine queries using Google. It is one of a suite of Google tools that track different types of activity, such as Google Scholar^{TM}, which records citations of papers, and Google Analytics^{TM}, which allows the owner of a website to track where and when people are viewing that site. As in any epidemiological study, it is important to take account of the denominator when interpreting counts. Google Trends does this by expressing the absolute number of searches relative to the total number of searches in each location and at each time. The number of searches for each term (e.g. "alcoholism") relative to total searches is referred to as the *query share*. This query share is then normalised to the highest volume of searches for that term over the time period being studied. This index ranges from 0 to 100, with 100 recorded on the date that saw the highest relative search volume activity for that term. Thus, a Google Trends^{TM} search index of 25 indicates that search activity for a particular term was 25% of that seen at the time when search activity was most intense. **Figure 2** illustrates this, and also how patterns can vary across countries. Thus, using searches for the term "antidepressants" (and equivalent in relevant languages) over the period 2004 to 2018, the highest intensities in three of the countries, Australia, France, and the United Kingdom, have been in 2017/18, with levels around 20 until the onset of the 2008 financial crisis. In contrast, the highest intensity in the USA was back in 2004.

Google Trends^{TM} offers a high level of geographical precision in developed countries, allowing for searches to be stratified at a national, regional, and city level. Trends in searches for different terms can be compared and multiple search terms can be combined, with a "+" sign, to identify those searching for terms in combination (e.g. "alcoholism + treatment").

Lastly, Google uses natural language processing methods and indexes web pages collected to classify search queries into one of 25 specific categories (including health) and over 300 sub-categories. Unlike other public health datasets, what makes Google Trends^{TM} data so novel is that it is collected and reported in real time and is publicly available.

Applications of Google Trends^{TM} in population health research

Initially, one of the best known applications of Google search engine query data in population health research was early detection of influenza epidemics. In 2009, collaborators at Google and the US Centers for Disease Control found that the relative frequency of searches for influenza-like illness correlated well with the percentage of physician visits for influenza in the United States, with a 1-day reporting lag (traditional CDC estimates have a 1-2 week reporting lag).[6] Since then, however, the uses of Google search engine data in population health research have grown substantially. Many examples involve early detection of other infectious outbreaks, such as Lyme disease[7], a selection of tropical diseases in India[8], syphilis[9], HIV[10], and Zika virus infections.[11]

Over time, Google search engine data has increasingly been utilized to understand health behaviours. For example, it has been shown that changes in the volume of suicide-related searches may provide an early warning of changing mental health risk,[12–14] although the association is strongest for suicides among younger people and middle-aged women, both groups more likely to take overdoses (and thus require information on how to do it) than among older men, among whom hanging is more common.[15] A related study examined the commonly held view that media coverage of celebrity suicides can either increase or decrease suicidal ideation, finding limited evidence for both, but only with the most prominent celebrities.[16]

Some other novel uses of Google Trends^{TM} have included monitoring interest in electronic cigarettes[17], abortions[18], and bariatric surgery[19], assessing the relationship between health-related searches and economic conditions (e.g. unemployment rates)[20–22], and tracking pharmaceutical utilisation and revenues[23] (**Boxes 1 and 2**). Several studies have also examined seasonality of events not easily identifiable from existing data sources.[24,25] A related application was seen in a study that quantified the number of days during that increased interest in smoking cessation lasted after a tax rise.[26] Information on what people search has even provided insights on societal attitudes, such as racism, which is less easily discernible in survey data (**Box 3**).[27]

The ability to combine search terms unlocks the potential for some especially imaginative approaches. For example, White et al.[28] demonstrated the ability to identify interactions between drugs from searches for their names in combination, something that would easily be missed using routine post-marketing surveillance.

Where we go from here with Google Trends^{TM}

While Google Trends^{TM} has been able to provide valuable insights into population health surveillance and behaviours, the data are subject to certain caveats. Among these limitations, the most obvious one is that all of the search data available through Google Trends^{TM} is anonymised and reflects those with internet access, potentially excluding vulnerable groups (e.g. elderly) or regions where internet uptake could be low (e.g. some parts of low- and middle-income countries). Researchers will not be able to know who is searching for health terms and what their intentions might be. While Google does use natural language processing methods to code health-related searches, this is not available for all countries and languages. Third, it is still not quite clear what search terms one should use when exploring a particular health behaviour (e.g. depression, alcoholism, etc.). Fourth, studies should be based on a clear conceptual model of behaviour, where appropriate, in which the searches can plausibly be linked to ideas and ultimately behaviour. This is not always the case.[29] Finally, it is important to be aware of the risk of reporting bias, with only those studies finding positive correlations being published, as has been suggested recently.[30]

Research using search engine activity is still in its infancy and there are some things that population health researchers and Google might do to maximise the value of this method.

The first relates to consistency of reporting. As Nuti et al.[31] note in their systematic review on Google Trends^{TM}, there is no defined consensus on how to document Google search engine queries in academic papers. For example, only 19% of the publications they identified had defined the search category used (e.g. health) and only 39% of papers provided an explicit search strategy. As a minimum, research utilising Google Trends^{TM} data should document the exact search terms inputted, the translations used when searching in other than English, category used, a downloadable spreadsheet of extracted search indices, and the date the analysis was performed.

The second relates to how health-related searches are collected and categorised, especially if a search term might be misconstrued as non-health related (e.g. "smoking" when juxtaposed with "chimney" or "gun"). This is an area where there is considerable scope for dialogue between population health researchers and Google, taking advantage of advances in artificial intelligence. Without any transparency on how algorithms for search terms are calculated, it will be difficult for researchers to know just how accurately search activity can model trends in search activity within a population.[32]

The third relates to ethical issues. Unlike research that uses postings on social media[33], Google Trends™ data are anonymised. However, it is conceivable that it may be possible to identify an individual living in a particular location who has a very rare disease, finding which other search terms were used in combination with that disease. Given advances in artificial intelligence, it will be important to monitor the situation for unintended consequences, such as those that have emerged with social media.[34]

Fourth, there is considerable scope for methodological development, drawing on new approaches to the analysis of search engine data from other fields. Thus, research on the propagation of memes, or ideas (e.g. marketing imagery, such as that employed by the manufacturers of products that may impact of health) has used concepts from infectious disease modelling.[35] One recent systematic review has identified forecasting as an area in particular need of development.[36]

Data from Google Trends™ and other search engine repositories will never replace traditional data collection methods for population health. However, with further refinements and constructive dialogue between researchers and Google, search engine data can offer a powerful, real-time tool to assess how population health and health behaviours are changing within society.

Box 1 – Tracking interest in electronic cigarettes: Evidence from Australia, Canada, United Kingdom, and the United States

In the past few years electronic cigarettes (e-cigarettes) have been marketed intensively.[28] Although the clinical evidence that they aid quitting is weak, social media abounds with claims that they are effective in this respect. To determine how consumer interest in these products has evolved, Ayers et al.[37] examined Google Trends^{TM} search activity for e-cigarettes against traditional smoking cessation products (e.g. nicotine replacement therapy) from July 2008 to February 2010. Their results showed that search activity for e-cigarettes rapidly surpassed that of any of the traditional smoking cessation products. For example, search volume for e-cigarettes was 300% and 160% higher than Chantix® or Champix® (varenicline) in the US and UK, respectively. Furthermore, search activity for e-cigarettes within the US was significantly higher for those states that had stronger tobacco control measures, as determined by the American Lung Association.

Box 2 – Association of abortion-related searches with abortion policies and availability: A global perspective

There are few, if any, medical procedures that provoke as much intense political debate as aborting a foetus. Moreover, as abortions remain illegal or heavily restricted in many jurisdictions, leading women to obtain them illegally or in other countries, accurate data on abortion provision is often unreliable and outdated. Because information identifying those undertaking internet searches is not in the public domain, data on volume of searches may provide a proxy for interest in obtaining an abortion. Based on this assumption, Reis and Brownstein[29] explored the relationships between Google Trends^{TM} search activity, local abortion rates, and local abortion policies. Both in the US and internationally, the study found search volume for abortions was inversely related to local abortion rates. However, search volume was also significantly higher in those regions of the US where barriers to an abortion were more stringent (e.g. mandatory parental notification for minors).

Box 3 – Uncovering racism

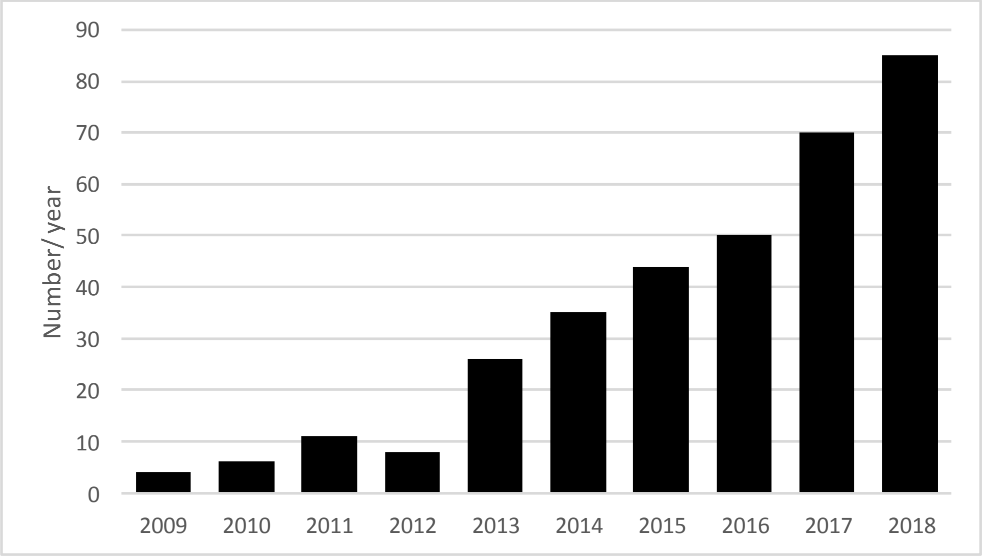
Social desirability bias exists where people answer questions in ways that they expect will be viewed favourably by others. This makes research on issues such as racism difficult. One study measured the proportion of Google searches for the “N-word” in 196 locations covering the USA.[27] They found a significant association with mortality among African Americans, even after adjusting for white mortality rates. As the authors noted, these findings are consistent with other literature on the adverse associations between racism and health.

Key Messages

- Google Trends^{TM} is the principle tool used to study trends and patterns of search engine queries—including health-related queries—using Google.
- Search engine data in public health has diverse applications in the literature, from tracking influenza outbreaks to monitoring interest in e-cigarettes.
- Google Trends^{TM} still has significant limitations that need to be addressed through dialogue between population health researchers and Google, particularly regarding how search engine queries are collected, organised, and coded.

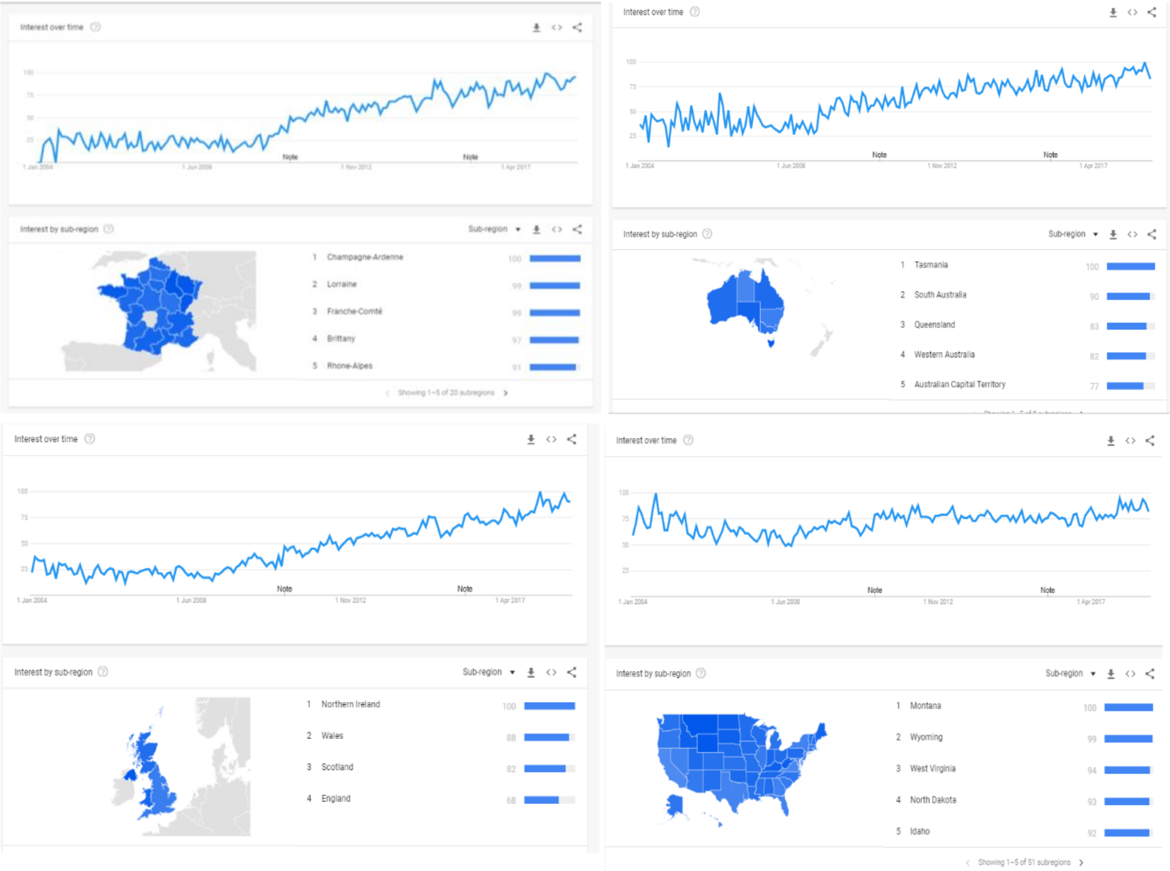
Figures

Figure 1: Research using Google Trends/ Insights, 2009 to 2018



Source: PubMed search for records with “Google Trends” or “Google Insights” in title or abstract

Figure 2: Search activity for anti-depressants in France, Australia, United Kingdom, and USA January 2004 to December 2018



1	Champagne-Ardenne	100
2	Lorraine	99
3	Franche-Comté	99
4	Brittany	97
5	Rhone-Alpes	91

‹ Showing 1–5 of 20 subregions ›

1	Tasmania	100
2	South Australia	90
3	Queensland	83
4	Western Australia	82
5	Australian Capital Territory	77

1	Northern Ireland	100
2	Wales	88
3	Scotland	82
4	England	68

1	Montana	100
2	Wyoming	99
3	West Virginia	94
4	North Dakota	93
5	Idaho	92

‹ Showing 1–5 of 51 subregions ›

Note: Search terms – France, antidepressants + antidépresseur; UK/ Australia – antidepressants; USA - antidepressants + antidepresivo (Spanish)

Source: Google Trends^{TM} searched 31st December 2018

References

1. Scott, M. Qwant Wants to Be Alternative to Google. *Bits Blog* (2014).
2. Search engine market share in the United States 2018 | Statistic. *Statista* Available at: https://www.statista.com/statistics/267161/market-share-of-search-engines-in-the-united-states/. (Accessed: 3rd January 2019)
3. Google Zeitgeist 2012: A Year in Search. Available at: http://www.google.co.uk/zeitgeist/2012/#the-world. (Accessed: 8th February 2015)
4. msnbc.com, J. W. More people search for health online. *msnbc.com* (2003). Available at: http://www.nbcnews.com/id/3077086/t/more-people-search-health-online/. (Accessed: 3rd January 2019)
5. NW, 1615 L. St, Washington, S. 800 & Inquiries, D. 20036 U.-419-4300 | M.-419-4349 | F.-419-4372 | M. Internet & Technology - Pew Research Center.
6. Ginsberg, J. *et al.* Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
7. Kapitány-Fövény, M. *et al.* Can Google Trends data improve forecasting of Lyme disease incidence? *Zoonoses Public Health* **0**,
8. Verma, M. *et al.* Google Search Trends Predicting Disease Outbreaks: An Analysis from India. *Healthc. Inform. Res.* **24**, 300–308 (2018).
9. Young, S. D., Torrone, E. A., Urata, J. & Aral, S. O. Using Search Engine Data as a Tool to Predict Syphilis. *Epidemiology* **29**, 574–578 (2018).
10. Young, S. D. & Zhang, Q. Using search engine big data for predicting new HIV diagnoses. *PLOS ONE* **13**, e0199527 (2018).
11. Morsy, S. *et al.* Prediction of Zika-confirmed cases in Brazil and Colombia using Google Trends. *Epidemiol. Infect.* **146**, 1625–1627 (2018).

12. McCarthy, M. J. Internet monitoring of suicide risk in the population. *J. Affect. Disord.* **122**, 277–279 (2010).
13. Gunn, J. F. & Lester, D. Using google searches on the internet to monitor suicidal behavior. *J. Affect. Disord.* **148**, 411–412 (2013).
14. Ayers, J. W., Althouse, B. M., Allem, J.-P., Rosenquist, J. N. & Ford, D. E. Seasonality in seeking mental health information on Google. *Am. J. Prev. Med.* **44**, 520–525 (2013).
15. Arora, V. S., Stuckler, D. & McKee, M. Tracking search engine queries for suicide in the United Kingdom, 2004-2013. *Public Health* **137**, 147–153 (2016).
16. Gunn, J. F., Goldstein, S. E. & Lester, D. The Impact of Widely Publicized Suicides on Search Trends: Using Google Trends to Test the Werther and Papageno Effects. *Arch. Suicide Res.* **0**, 1–24 (2018).
17. McKee, M. E-cigarettes and the marketing push that surprised everyone. *BMJ* **347**, f5780–f5780 (2013).
18. Reis, B. Y. & Brownstein, J. S. Measuring the impact of health policies using Internet search patterns: the case of abortion. *BMC Public Health* **10**, 514 (2010).
19. Rahiri, J.-L. *et al.* Using Google Trends to explore the New Zealand public’s interest in bariatric surgery. *ANZ J. Surg.* **88**, 1274–1278 (2018).
20. Althouse, B. M., Allem, J.-P., Childers, M. A., Dredze, M. & Ayers, J. W. Population health concerns during the United States’ Great Recession. *Am. J. Prev. Med.* **46**, 166–170 (2014).
21. Tefft, N. Insights on unemployment, unemployment insurance, and mental health. *J. Health Econ.* **30**, 258–264 (2011).
22. Frijters, P., Johnston, D. W., Lordan, G. & Shields, M. A. Exploring the relationship between macroeconomic conditions and problem drinking as captured by Google searches in the U.S. *Soc. Sci. Med. 1982* **84**, 61–68 (2013).

23. Nathaniel M. Schuster, B. S., Mary A. M. Rogers, P. & and Laurence F. McMahon Jr, M. D. Using Search Engine Query Data to Track Pharmaceutical Utilization: A Study of Statins. *Am. J. Manag. Care* **16**, (2010).

24. Association of Search Engine Queries for Chest Pain With Coronary Heart Disease Epidemiology. | Cardiology | JAMA Cardiology | JAMA Network. Available at: https://jamanetwork-com.ezp-prod1.hul.harvard.edu/journals/jamacardiology/fullarticle/2706608. (Accessed: 3rd January 2019)

25. Seasonal and Geographic Patterns in Seeking Cardiovascular Health Information: An Analysis of the Online Search Trends - ScienceDirect. Available at: https://www-sciencedirect-com.ezp-prod1.hul.harvard.edu/science/article/pii/S0025619618305779?via%3Dihub. (Accessed: 3rd January 2019)

26. Tabuchi, T., Fukui, K. & Gallus, S. Tobacco Price Increases and Population Interest in Smoking Cessation in Japan Between 2004 and 2016: A Google Trends Analysis. *Nicotine Tob. Res.* doi:10.1093/ntr/nty020

27. Chae, D. H. *et al.* Association between an Internet-Based Measure of Area Racism and Black Mortality. *PLOS ONE* **10**, e0122963 (2015).

28. White, R. W., Tatonetti, N. P., Shah, N. H., Altman, R. B. & Horvitz, E. Web-scale pharmacovigilance: listening to signals from the crowd. *J. Am. Med. Inform. Assoc. JAMIA* **20**, 404–408 (2013).

29. Low validity of Google Trends for behavioral forecasting of national suicide rates. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0183149. (Accessed: 3rd January 2019)

30. Cervellin, G., Comelli, I. & Lippi, G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. *J. Epidemiol. Glob. Health* **7**, 185–189 (2017).

31. Nuti, S. V. *et al.* The Use of Google Trends in Health Care Research: A Systematic Review. *PLoS ONE* **9**, e109583 (2014).

32. Butler, D. When Google got flu wrong. *Nature* **494**, 155–156 (2013).

33. McKee, R. Ethical issues in using social media for health and health care research. *Health Policy* **110**, 298–301 (2013).

34. McKee, M. & Stuckler, D. How the Internet Risks Widening Health Inequalities. *Am. J. Public Health* **108**, 1178–1179 (2018).

35. Wang, L. & Wood, B. C. An epidemiological approach to model the viral propagation of memes. *Appl. Math. Model.* **35**, 5442–5447 (2011).

36. Mavragani, A., Ochoa, G. & Tsagarakis, K. P. Assessing the Methods, Tools, and Statistical Approaches in Google Trends Research: Systematic Review. *J. Med. Internet Res.* **20**, e270 (2018).

37. Ayers, J. W., Ribisl, K. M. & Brownstein, J. S. Tracking the Rise in Popularity of Electronic Nicotine Delivery Systems (Electronic Cigarettes) Using Search Query Surveillance. *Am. J. Prev. Med.* **40**, 448–453 (2011).