Using Google Trends in Infodemiological studies in Sexual and Reproductive Health Research

Mona Larki (PhD)[1], Elham Manouchehri (PhD)[2], Robab Latifnejad Roudsari (PhD)[3, 4*]

[1] PhD of Reproductive Health, Nursing and Midwifery Care Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[2] PhD of Reproductive Health, Department of Midwifery, Faculty of Nursing and Midwifery, Mashhad Medical Sciences, Islamic Azad University, Mashhad, Iran
[3] Professor, Nursing and Midwifery Care Research Center, Mashhad University of Medical Sciences, Mashhad, Iran
[4] Department of Midwifery, School of Nursing and Midwifery, Mashhad University of Medical Sciences, Mashhad, Iran

Internet search activities may provide one form of Big Data, which could create great and valuable insights into disease patterns and populations' health-related behaviours (1). The meaning of big data in health sciences can be quantitative (volume of data) or qualitative (complexity of data) (2). New tools are emerging to make health care research possible in the Big Data era (1). One of these tools that provide the basis for internet search data is Google Trends (GTs) or Search Engine Query Data (SEQD), a publicly accessible resource of online Google search trafficking data (https://trends.google.co.jp /trends) that includes both real-time and archival data. It enables users to observe variations in periods associated with the general public online interest in specific keywords from around 2004. GTs analyses a portion of the three billion daily Google Search searches (1). In order to inform public health and policy, GTs has been used to track changes in web-based interest over time, evaluate correlations between search terms and other data sources, and predict disease susceptibility and incidences (3).

Reasons exist for GTs surge to prominence in the field of big data investigation and applications. Clients' wants, needs, and demands could be promptly reflected through GTs analysis. Therefore, the information-seeking patterns of users can be easily investigated. It is a user-friendly and rich platform that not only collects data but also provides some tools for comparing different topics (4). New research shows that Google search data can be used to keep track of a wide range of social and biological issues when more reliable or up-to-date data is not available (5).

GTs is a popular data source for "infodemiology", which according to Eysenbach (2000), is an area of research focused on scanning the internet for user-contributed health-related content, with the ultimate goal of improving public health. It is used to study the patterns of disease outbreaks, health topics, and challenges (4, 6, 7). It necessitates a number of special methodologies for consumer and public health informatics to describe and analyze health information and communication trends in electronic media, and quantify the epidemiology of information (7). GTs can also be used for infoveillance, which is the "longitudinal observation of infodemiology data for surveillance and trend analysis" (5). The monitoring and prediction of health-related topics are thus made possible by information from anonymous internet searches (5, 8).

By combining and comparing GTs data with survey results about the availability of services and healthcare facilities in different areas, demand-supply gaps can be found. This could be especially helpful in cases like the COVID-19

* *Corresponding author:* Robab Latifnejad Roudsari, Professor, Nursing and Midwifery Care Research Center, Mashhad University of Medical Sciences, Mashhad, Iran. Tel: 05138591511; Email: rlatifnejad@yahoo.com, LatifnejadR@mums.ac.ir

pandemic, where it is difficult to collect data on people (9). Although web mining is an interesting approach, it is an alternative for the efforts of public health care institutions and health researchers to collect "real-life" epidemiologic information (10). As research with this methodology framework expands, future studies should address substantial limitations that must be resolved through interaction between health researchers and Google, particularly in the collection and organization of search terms (11). Because of the activities of millions of users, Google and the internet as a whole are continually changing. To further comprehend how these alterations develop over time, researchers need more information. To make sure they are studying real trends and not temporary patterns, scientists must reproduce their results using these sources of data throughout time and utilizing additional data resources (12).

In the field of sexual and reproductive health (SRH), internet search is done for various issues and topics. This is undeniable that most of the women, especially from developed countries use search engines to find SRH information online at some point (13). Predictable behaviors and information-seeking during pregnancy for some disorders like morning sickness (5) or for decision-making about mode of delivery (14) makes Google data search a great tool for predicting what will happen (5). Although, there are concerns that low-quality and out-of-date content on the internet may induce unnecessary cesarean section (11). Information seeking in clinical settings about issues such as sexual attitudes and behaviors (15, 16), abortion and pregnancy termination in teen pregnancies (17) is limited due to legislative limitations, stigmatization and dread of potential legal outcomes (18). Additionally, women may use the internet more to get important information about family planning, especially during the pandemics and lock down that followed (9), and for being informed about how to report domestic violence (19). Thus, due to the widespread use of the internet, experts in reproductive health should take part in online discussions to spread accurate information and point people in the right direction of reliable resources (11). Therefore, women's SRH issues

around the world are highlighted as a potential application of internet search data using GTs tool (11, 13, 16, 17, 20).

Even though there has been a lot of GTs research in the field of health care in recent years, there is still no direction or framework of accepted standards for the appropriate use of this tool and the relevant studies lack a well-defined methodology framework. But to apply it reliably as a research tool, it needs to be more obvious. This makes it more useful for health-care studies in general and increases the trustworthiness of the generated data (1, 10).

GTs requires an intuitive approach, it may be implemented more quickly and its results are available without delay compared to other more specialist methods like sentiment analysis (SA) (21). To establish a strong methodological framework for using GTs data, the selection of the proper keyword(s), region(s), time period, and category are the major factors that must be taken into account, which is essential for ensuring the quality and validity of the results (22).

Choosing the right keyword(s): GTs does not differentiate between capital and lowercase letters, but it does take into account accents, plural and singular forms, and spelling errors. Therefore, regardless of the keyword or keyword combination used, some portions of the relevant query will not be analysed further (22).

Region selection: The choice of the geographic area for which query data will be retrieved is another salient issue. Data can be downloaded for one or more terms with global or national online interest at the first level of categorisation. There is a list of all the countries, and in the majority of them, one can research interests in smaller regions. Users can choose the geographic region to research a city, a country, or the entire world, and data is accessible for every country in the world (3).

Period Selection: One of the most frequent errors in GTs research is the choice of the analysed time period. The user can select a period of time to research, which can be divided into months or days and ranges, for instance, from January 2004 to the present (22). The timeframe at which GTs data are gathered is essential for the validity of the results because

the data are normalized across the chosen period. The basic rule is that the time period chosen for Google data should perfectly match the time period for which official statistics are available and will be assessed(23).

Search Categories: The chosen keyword (s) can be examined in relation to a chosen category while researching an internet interest. When the same word is used or may be connected to various meanings or events, this feature is crucial for removing noisy data (22).

Another methodological issue is that relative search volumes (RSVs) of one search word can be compared over periods of time and geographical regions, as well as the RSVs of up to five other search terms. The user can further narrow their search by selecting from 25 distinct topic areas, each with several subcategories, for a total of >300 options (1). To find out if there was a date dependency, RSVs of a particular query over a certain time period should be downloaded on multiple days. Based on Google, RSVs are made by dividing each data point by the total number of searches it represents for the location and time period. Otherwise, the top-ranked locations would always be those with the largest search volume. The data is then scaled from 0 to 100 based on the frequency of searches for each topic versus all searches for all topics (24). RSVs must be carefully interpreted and analyzed in the right way because, in addition to being affected by social media, they can also be affected by random changes and abnormalities (21).

The main advantage of GTs is that they employ both visible and invisible user choices. Therefore, users can obtain data that would be difficult or impossible to obtain. Also, because the data is accessible in real-time, it eliminates problems associated with conventional, time-consuming survey methodologies. Due to the confidentiality of internet searches, it is possible to analyze and predict sensitive diseases and subjects such as AIDS, mental illnesses, and suicides as well as using illegal narcotics(25).

However, there are various restrictions on how GTs data can be utilized. First, despite the obvious promise of Google data in epidemiology and disease surveillance, online search traffic data has not always been able to accurately predict the spread of disease, as it was the case with Google Flu Trends (3). This may be partly because when using GTs to do research, the sample size is not clear and cannot be shown to be representative. Also, online searches do not give reliable results in places where people do not have access to the internet or the freedom to speak their opinions (2). Thirdly, the choice of the keyword(s) is crucial to assuring the veracity of the results. In particular, careful assessment should be undertaken to guarantee that news coverage and unexpected events do not damage the validity of the results. Additional demographic characteristics such as age and gender cannot be included in the study because the sample size is unknown. As this is a relatively new field of study, there is no standardized structure for its reports. When GTs is used to answer the same question, choosing different terms can result in various findings and conclusions. It is important to explain why these terms were chosen so that the reader can better fully comprehend the research methodologies and the face validity of the study (21). Also, some terms mean the same thing while others have multiple meanings and abbreviations (3). Fourth, Google uses natural language processing algorithms to code health-related searches, but not everywhere or in every language (4).

It should be noted that the quality factors, which are specified and applied to GTs data include accuracy, completeness, consistency, and validity. The accuracy of the data is seen to be of particular importance; since, if left unadjusted, a low level of accuracy could be a substantial source of bias (26). Also, awareness of the risk of reporting bias is important because it has recently been suggested that only surveys reporting positive correlations be published (10). But even with limitations, GT is still a useful source of information for social and economic studies (26). Although GTs can bring several insights and research opportunities and investigate health challenges, there are problems with the documentation of the methodology. Thus, to enhance its validity, the documentation is important because inadequate documentation of procedures prevents replication of the results. In addition, increased transparency can enhance

its reliability as a research instrument. This documentation would enable other scholars to assess the consistency over time of the results obtained by GTs in response to a specific query (1). Data documentation is also crucial to make sure that the results can be reproduced, which is a key part of good science (1). Obviously, it is a well-known fact that scientific research that cannot be replicated makes it much less useful and reliable (27). To ensure the absolute trustworthiness of a GTs dataset, it is crucial for future studies that researchers collect queries data from multiple consecutive days and analyze them utilizing their RSVs average rather than daily RSVs, minimizing standard errors until a predetermined confidence threshold is achieved (24).

In conclusion, GTs could be deemed as a free and easily accessible means to access large amounts of search data to derive meaningful insights about population health behaviors. However, to be employed reliably as a research instrument would require being more transparent, which is essential for guaranteeing the value and validity of both the results obtained and its general usefulness to health care research in general and particularly to sexual and reproductive health care research. Although GTs and other search engine databases will not ever displace traditional methods of obtaining information about SRH, if GTs data is properly validated, it could become an important and complementary source of additional data for researchers and policymakers studying the unique topics and challenges of SRH at the local, national, and global levels.

Conflicts of interest

Authors declared no conflict of interest.

References

1. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: a systematic review. PloS one. 2014; 9(10): e109583.
2. Shilo S, Rossman H, Segal E. Axes of a revolution: challenges and promises of big data in healthcare. Nature Medicine. 2020; 26(1): 29-38.
3. Zepecki A, Guendelman S, DeNero J, Prata N. Using application programming interfaces to access Google data for health research: Protocol for a methodological framework. JMIR Research Protocols. 2020; 9(7): e16543.
4. Jun S-P, Yoo HS, Choi S. Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technological forecasting and social change. 2018; 130 (C): 69-87.
5. Wilde J, Chen W, Lohmann S. COVID-19 and the future of US fertility: what can we learn from Google. 2020.
6. Arora VS, McKee M, Stuckler D. Google Trends: Opportunities and limitations in health and health policy research. Health Policy. 2019; 123(3): 338-341.
7. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. Journal of Medical Internet Research. 2009; 11(1): e1157.
8. Alshahrani R, Babour A. An infodemiology and infoveillance study on COVID-19: analysis of twitter and google trends. Sustainability. 2021; 13(15): 8528.
9. Dey A, Dehingia N, Raj A. Using Google trends data to assess reproductive health needs in Nigeria during COVID-19. Big Data and Gender in the Age of Covid. 2020.
10. Cervellin G, Comelli I, Lippi G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. Journal of Epidemiology and Global Health. 2017; 7(3): 185-189.
11. Kamiński M, Łoniewski I, Łoniewska B. 'Dr. Google, is caesarean section good for me?' - the global internet searches associated with mode of birth methods: Retrospective analysis of Google trends data. Midwifery. 2020; 89: 102787.
12. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science. 2014; 343(6176): 1203-1205.
13. Cheng RfJ, Fisher AC, Nicholson SC. Interest in Home Birth During the COVID-19 Pandemic: Analysis of Google Trends Data. Journal of Midwifery & Women's Health. 2022; 67(4): 427-434.
14. Kamiński M, Łoniewski I, Łoniewska B. 'Dr. Google, is caesarean section good for me?'–the global internet searches associated with mode of birth methods: Retrospective analysis of Google trends data. Midwifery. 2020; 89: 102787.
15. MacInnis CC, Hodson G. Do American states with more religious or conservative populations search more for sexual content on Google? Archives of Sexual Behavior. 2015;

44(1): 137-147.

16. Ojala J, Zagheni E, Billari F, Weber I. Fertility and Its Meaning: Evidence from Search Behavior. Proceedings of the International AAAI Conference on Web and Social Media. 2017; 11(1): 640-643. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14915

17. Reidpath DD, Allotey P. Predicting US state teenage birth rates using search engine query data on pregnancy termination and prevention. Journal of Global Health Reports. 2018; 2(9): 1-4.

18. Jerman J, Onda T, Jones RK. What are people looking for when they Google "self-abortion"? Contraception. 2018; 97(6): 510-514.

19. Guatimosim RF, Teles ALS, Loureiro FF, da Silva AG, de Miranda DM, Malloy-Diniz LF. What do we know about violence against women in pandemic times? Insights based on search trends. medRxiv. 2021.

20. Dey A, Dehingia N, Raj A. Understanding patterns of unplanned pregnancies and abortions in India during the COVID-19 pandemic using Google Trends data. [Internet]. 2021 [cited 30 sep 2022]. Available from: https://emerge.ucsd.edu/understanding-patterns-of-unplanned-pregnancies-and-abortions-in-india-during-the-covid-19-pandemic-using-google-trends-data/

21. Rovetta A, Castaldo L. A new infodemiological approach through Google Trends: longitudinal analysis of COVID-19 scientific and infodemic names in Italy. BMC medical research methodology. 2022; 22(1): 1-14.

22. Mavragani A, Ochoa G. Google Trends in infodemiology and infoveillance: methodology framework. JMIR public health and surveillance. 2019; 5(2): e13439.

23. How to Use Google Trends for Keyword Research: 7 Effective Ways. [Internet]. 2019 [cited 30 sep 2022]. Available from: https://ahrefs.com/blog/how-to-use-google-trends-for-keyword-research/.

24. Rovetta A. Reliability of Google Trends: Analysis of the limits and potential of web infoveillance during COVID-19 pandemic and for future research. Res. Metr. Anal. 6:670226.

25. Jun S-P, Yoo HS, Choi S. Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. Technological forecasting and social change. 2018; 130: 69-87.

26. Cebrián E, Domenech J. Is Google Trends a quality data source? Applied Economics Letters. 2022: 1-5.

27. Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJ, Fiedler K, et al. Recommendations for increasing replicability in psychology. 2013; 27: 108–119.