

# The Use of Google Trends for Analysis of Interest for Information about Cancer in Brazil: Theoretical-Methodological Aspects

doi: <https://doi.org/10.32635/2176-9745.RBC.2023v69n2.3847>

*O Uso do Google Trends para Análise de Interesse por Informações sobre o Câncer no Brasil: Aspectos Teórico-Metodológicos*  
El Uso de Google Trends para el Análisis de Interés en Información sobre el Cáncer en el Brasil: Aspectos Teórico-Metodológicos

Fernando Lopes Tavares de Lima<sup>1</sup>

## INTRODUCTION

In Brazil, 704,000 new cases of cancer are estimated for each year of the 2023-2025<sup>1</sup> triennium. This scenario reaffirms cancer as a growing public health problem in the country. The current National Policy for Cancer Prevention and Control (PNPCC)<sup>2</sup> establishes the stimulus for the formulation of communication strategies with the population, aiming to disseminate and expand knowledge about cancer, its risk factors and prevention guidelines.

Such a stimulus should accompany practical actions that expand society's knowledge on preventing and detecting cancer early<sup>3</sup>. However, given the difficulties of accessing these actions, the Internet has been used and, in this environment, the *Google* search engine is a privileged space, considered an important source of health information<sup>4</sup>. The widespread use of this tool has generated a huge database of the population's interest in any topic, including cancer.

Researchers have used *Google Trends* (GT) as a tool to access this data and identify population interests in health information<sup>5</sup>. Thus, Internet-based behavior is analyzed to examine actual human behavior<sup>6</sup>.

Despite the importance of research with the TG, there are concerns about the methodological robustness of the articles. In 2014, it was identified that only 7% of the articles provided the necessary information to reproduce the search, and only 34% documented the date on which the search was performed<sup>7</sup>.

Thus, given the importance of this tool, the objective of this article is to present a theoretical-methodological orientation on its use, as well as its indications and limitations in the scope of cancer prevention and control in Brazil.

## DEVELOPMENT

### THE GOOGLE TRENDS

GT is a free online portal created in 2006 by *Google Inc.* that presents the relative frequency with which a given term is searched in the *Google* search engine worldwide<sup>8</sup>. Thus, based on the actual queries of users, the TG provides information on what was and what is trending in the interests of Internet users<sup>6</sup>.

Due to its rapid processing, data are available in real time<sup>5</sup>, and its use has been associated with its usefulness in promptly detecting a phenomenon, its representativeness, and its high potential for application in the most diverse areas<sup>9</sup>.

In general, research using the WG aims at four objectives: measuring interest in a topic, detecting its variations or seasonality, analyzing the relationships of this variation with specific events, and predicting health-related events<sup>6</sup>.

### USING GOOGLE TRENDS

To use GT, go to <https://trends.google.com/> and enter a search strategy. From the presentation of the results, the set of tools to improve this search is available: (1) search strategy, (2) geographic area, (3) period, (4) category, and (5) search modality. Another way to access the tools is to click on the "Explore" tab on the home page.

The construction of a search strategy should consider no differentiation between uppercase and lowercase letters, but there is concerns about accents, plural forms, and spelling errors<sup>8</sup>. An example is the search for the term "prostate cancer" with and without accent. Despite the correct form of writing with accent marks in Brazil, there is a greater volume of searches for expressions without accent marks (Figure 1A). This exemplifies the importance

<sup>1</sup>Instituto Nacional de Câncer (INCA), Coordenação de Prevenção e Vigilância (Conprev), Divisão de Pesquisa Populacional. Rio de Janeiro (RJ), Brazil. E-mail: [flima@inca.gov.br](mailto:flima@inca.gov.br). Orcid iD: <https://orcid.org/0000-0002-8618-7608>

**Corresponding author:** Fernando Lopes Tavares de Lima. Rua Marquês de Pombal, 125, 7º andar – Centro. Rio de Janeiro (RJ), Brazil. CEP 20230-240. E-mail: [flima@inca.gov.br](mailto:flima@inca.gov.br)



of the researcher analyzing the different search possibilities in their language and building a strategy that does not lose data.

The search strategy can be composed of terms and expressions. The use of quotation marks in expressions limits the results to this exact form of writing<sup>7</sup>. Figure 1B exemplifies a slight variation for a search for “prostate cancer” with and without the quotation marks.

Searching some terms in GT automatically opens search options for specific topics. Topics are a group of synonymous terms already categorized in any language, which increases the power of search<sup>8</sup>. An example is presented in Figure 1C, which compares the search for “prostate cancer” as a search term and as a topic.

Another important strategy is the aggregate search of terms or expressions. To this end, the “+” feature is used. This aggregates different terms within the same search strategy group (limited to five simultaneous search groups of 25 terms each). Thus, variations of the term can be included, as well as common spelling errors<sup>5</sup>.

A Geographic Region of coverage must be delimited. You can select the whole world or a specific country. In addition, some countries have data by sub-regions<sup>5,7</sup>, such as the Federative Units in Brazil.

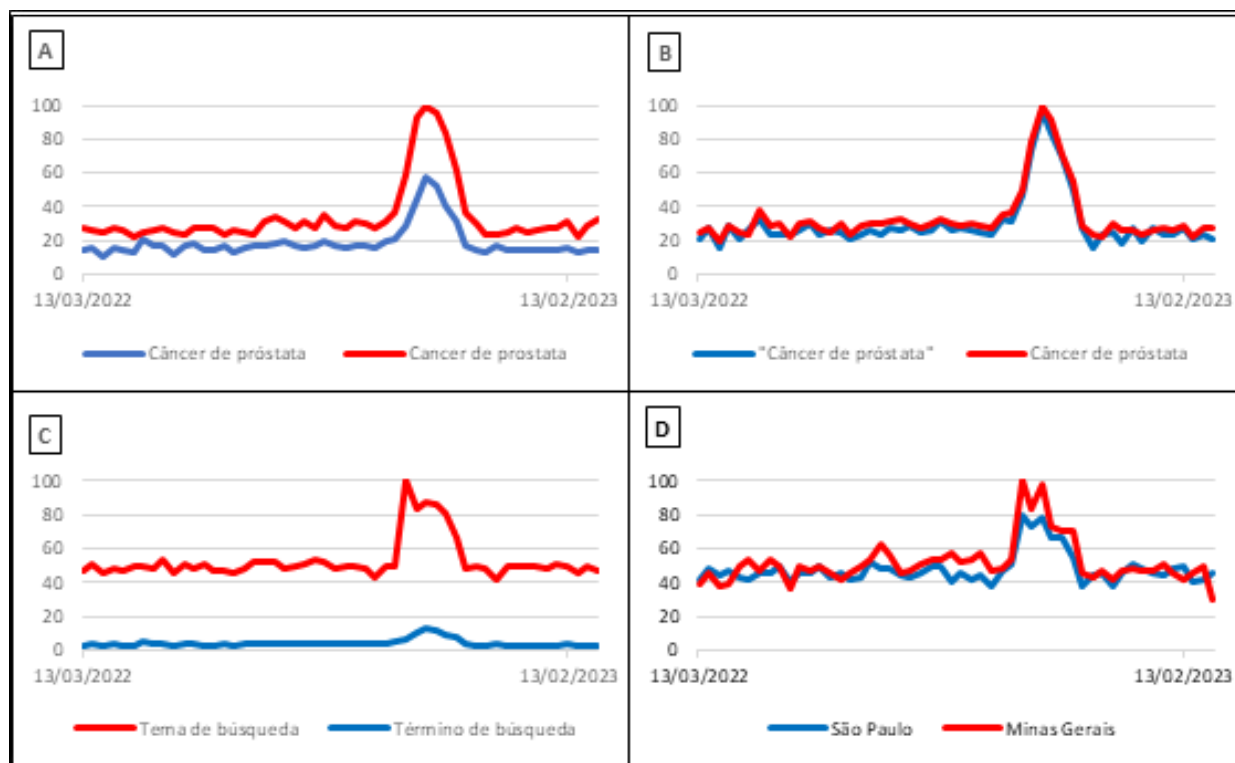
Figure 1D compares two regions, São Paulo and Minas Gerais, for the topic “Prostate cancer” in the last year. The higher value in Minas Gerais does not mean a higher absolute value, as the measure is relative<sup>8</sup>.

The WG offers two types of data. The first is classified as “real time” and corresponds to the last seven days. The second is a separate sample from the first that goes from 2004 to up to 72 hours before the survey<sup>8</sup>.

The researcher can choose for a period already configured or customize it. Regardless of the choice, it is important to note that the data is normalized for the selected period. Thus, the search should be composed exactly by the period of interest<sup>7</sup>.

The search can be performed based on a category. This feature is important to eliminate data noise, especially in cases where the word has different meanings. For example, the word “cancer” can mean a disease or the sign of the zodiac. The selection of the category “health” can delimit the results (effect similar to the choice of search by topic). It should be noted that this affects all groups of search terms.

As a search modality, GT offers the options: “Web Search,” “Image Search,” “News Search,” “Google Shopping,” and “YouTube Search.” The “Web Search”



**Figure 1.** Examples of using different tools to build the search strategy in Google Trends. Image **A**: difference in the use of the accent in the expression “prostate cancer” in Brazil. Image **B**: difference in the use of quotation marks in the expression “prostate cancer” in Brazil. Image **C**: difference in the selection of topic or search term for the expression “prostate cancer.” Image **D**: difference in the search volume for “prostate cancer” in two Federation Units in Brazil

**Source:** Google Trends<sup>8</sup>; accessed in 4/3/2023.

option refers to the Google search engine and is the object of this article. However, depending on the research objective, another selection can be made<sup>5</sup>. Selecting the search modality affects all groups of search terms.

**SEARCH RESULTS**

The result of the search in the TG is the relative search volume (RV), scaled from zero to 100. The zero value indicates very low search volumes, and the value of 100 represents the peak of interest<sup>8</sup>. This index is a normalized value for a given location and period, allowing comparisons<sup>8</sup>.

The results in the WG online platform are presented in four panels: *Interest over time*, *interest by sub-region*, *Related subjects*, and *Related surveys*.

**Interest over time comparison**

This panel presents a line chart, where the vertical axis is the VPR value, and the horizontal is the time scaled

depending on the selected period (in minutes, hours, days, weeks, and months).

Figure 2A reproduces the *Interest panel over time*, which shows the difference between the interest of Brazilian and Portuguese Internet users in the topic “Prostate cancer” in the last year.

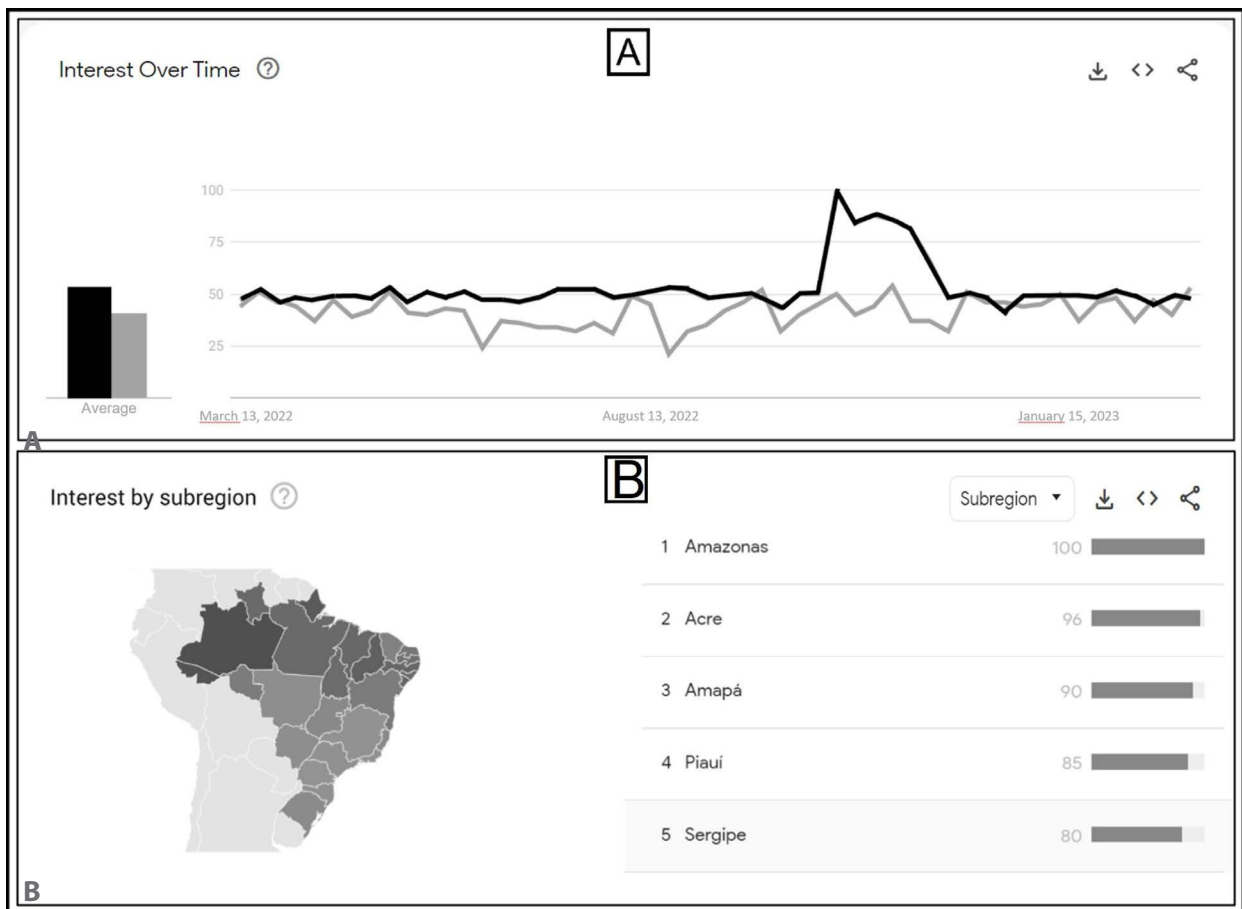
**Interest by sub-region**

The second panel presents a map graph dividing the selected region into its subregions, with darker colors where the VPR is larger. In addition to the map chart, a list of the sub-regions sorted according to the popularity of the term<sup>8</sup>.

Figure 2B exemplifies the *interest by sub-region* panel in a search for Brazilian interest in “Cervical cancer” last year.

**Related Matters**

In this panel, topics that were also searched by users who searched for the term of interest are presented. It is



**Figure 2.** Reproduction of *Interest panels over time* and *Interest by Google Trends sub-region*. Image A: *Interest panel over time* of Brazilian (darker tone) and Portuguese (lighter tone) Internet users for the topic “prostate cancer” in the last year. Image B: *Interest panel by sub-region* of Brazilian Internet users for the topic “cervical cancer” in the last year, in which the darker images represent higher values

Source: Google Trends<sup>8</sup>, accessed in 4/3/2023.

an important tool for *insights*. This presentation is done on two metrics: key and rising. The main ones are the most researched related subjects for the entire historical series. The subjects on the rise are the ones that have had a sudden increase in frequency. Figure 3A and Figure 3B present the two-panel view options, respectively.

### Related Research

The last panel resembles the previous one in presentation and function but presents the terms that were also searched by users who searched for the topic of interest. Figure 3C (main) and Figure 3D (on the rise) exemplify the research related to the search of Brazilians for the topic “Prostate cancer” in the last five years.

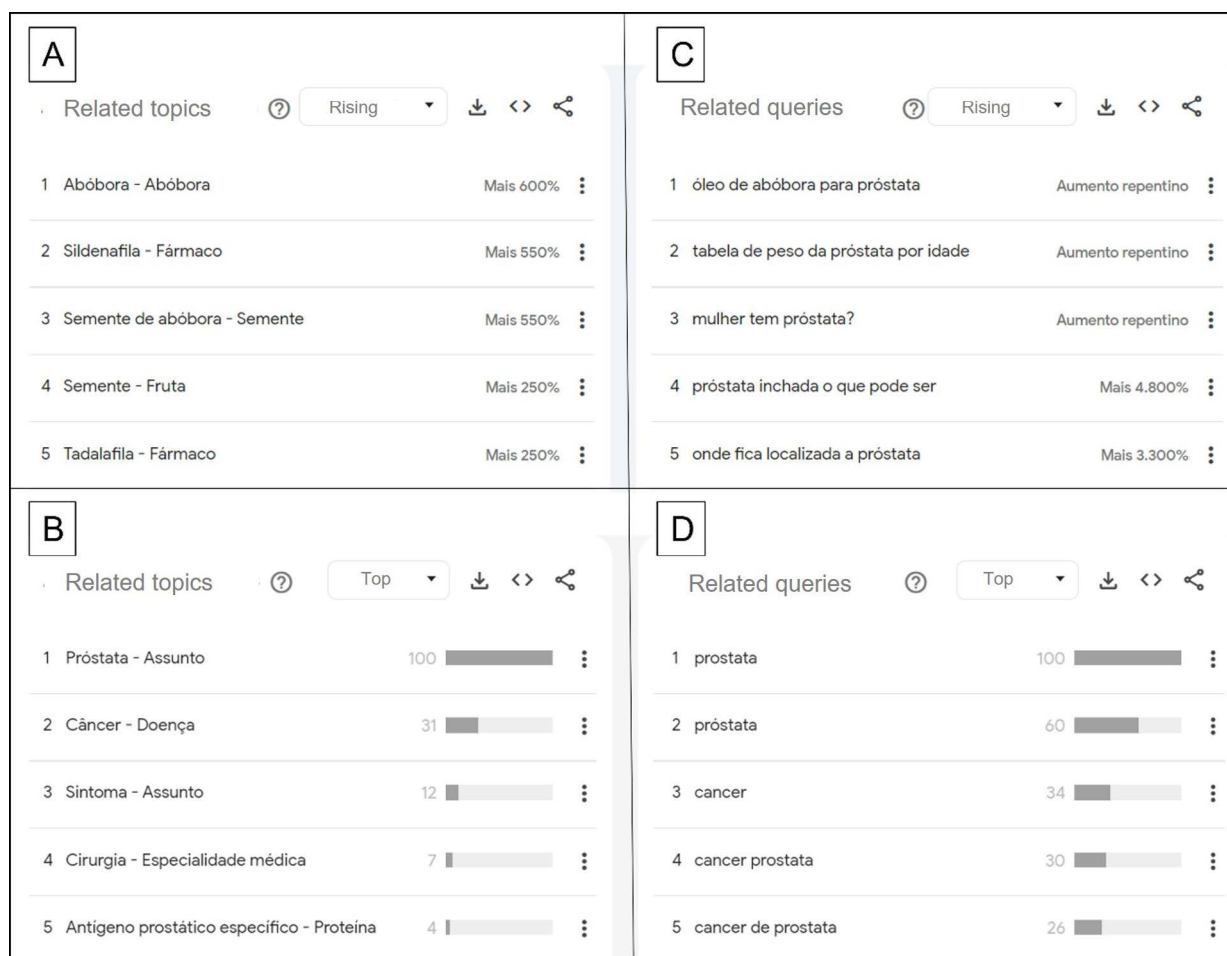
The four panels allow data export in a “CSV” file, which allows analysis in other *software*. In addition, the WG also allows the incorporation of panels on other *websites* or their shares on some social networks<sup>7,8</sup>.

Two tools can be used to monitor the values of interest. The first is by “enrolling” in the GT itself, selecting the terms and a region of interest. Thus, the user receives periodic updates by email<sup>5</sup>.

The second way is the use of the GT *Application Programming Interface* (API); however, programming knowledge is necessary. An example is Python’s *Pytrends* library. Although more complex, it allows for greater personalification of monitoring.

Data analysis in research with the TG is done through a quantitative approach. The choice of statistical tests varies according to the objectives of the study. Due to the various possibilities and limits of space, this process will not be detailed, but the review by Mavragani et al.<sup>6</sup> is suggested.

Four limitations to the use of TG data are commonly debated. The first is that the results presented are a sample of unknown size. The second is that regions with low Internet penetration or less use of the *Google* search engine



**Figure 3.** Reproduction of the subjects and related research of Google Trends, presenting the search for the interest of Brazilians in the topic “Prostate cancer” in the last five years

**Source:** Google Trends<sup>8</sup>; accessed in 4/3/2023.

may offer limited interest data. The third is the lack of absolute data, which reduces the possibilities of analysis. The fourth occurs when there is difficulty in selecting search terms, which compromises the exclusion of noise<sup>5,9</sup>.

Despite these aspects, studies empirically suggest that the results are valuable and valid in exploring and predicting behavior and are correlated with real data<sup>5</sup>.

## CONCLUSION

The GT allows access to a sample of the searches conducted by real users in the *Google* search engine. These data provide valuable insights into populations' health-related behavior and interest trends. The Brazilian data allows for a comparison between the different Federation Units and municipalities.

To improve the quality and reproducibility of the studies, paying attention to the criteria described is essential. And also details the search strategies, the region, and the period of interest, in addition to the categories and search locations. The date on which the search was performed and the statistical tests used must also be presented.

Considering that the GT presents results in real time and enables permanent monitoring, this portal is an important analysis tool for public health decisions, including the field of cancer prevention and control in the country.

## CONTRIBUTION

Fernando Lopes Tavares de Lima participated in all stages of the construction of the article, from its conception to the approval of the final version to be published.

## DECLARATION OF CONFLICT OF INTERESTS

There is no conflict of interest to declare.

## FUNDING SOURCES

None.

## REFERENCES

1. Santos MO, Lima FCS, Martins LFL, et al. Estimativa de incidência de câncer no Brasil, 2023-2025. *Rev Bras Cancerol.* 2023;69(1):e-213700. doi: <https://doi.org/10.32635/2176-9745.RBC.2023v69n1.3700>
2. Ministério da Saúde (BR), Gabinete do Ministro. Portaria nº 874, de 16 de maio de 2013. Institui a Política Nacional para a Prevenção e Controle do Câncer na Rede de Atenção à Saúde das Pessoas com Doenças Crônicas no âmbito do Sistema Único de Saúde (SUS) [Internet]. *Diário Oficial da União, Brasília, DF.* 2013 maio 17 [acesso 2023 mar 10]; Seção 1:129. Disponível em: [http://bvsmms.saude.gov.br/bvsm/saudelegis/gm/2013/prt0874\\_16\\_05\\_2013.html](http://bvsmms.saude.gov.br/bvsm/saudelegis/gm/2013/prt0874_16_05_2013.html)
3. Assis M. Comunicação em saúde na prevenção e detecção precoce do câncer: em busca de práticas mais dialógicas e inclusivas. *Rev Bras Cancerol.* 2023;69(1):e-032879. doi: <https://doi.org/10.32635/2176-9745.RBC.2023v69n1.2879>
4. Fahy E, Hardikar R, Fox A, et al. Quality of patient health information on the Internet: reviewing a complex and evolving landscape. *Australas Med J.* 2014;7(1):24-8. doi: <https://doi.org/10.4066/AMJ.2014.1900>
5. Mavragani A, Ochoa G. Google Trends in infodemiology and infoveillance: methodology framework. *JMIR Public Health Surveill.* 2019;5(2):e13439. doi: <https://doi.org/10.2196/13439>
6. Mavragani A, Ochoa G, Tsagarakis KP. Assessing the methods, tools, and statistical approaches in google trends research: systematic review. *J Med Internet Res.* 2018;20(11):e270. doi: <https://doi.org/10.2196/jmir.9366>
7. Nuti SV, Wayda B, Ranasinghe I, et al. The use of Google Trends in health care research: a systematic review. *PLoS One.* 2014;9(10):e109583. doi: <https://doi.org/10.1371/journal.pone.0109583>
8. Google Trends [Internet]. California: Google; c2023 [acesso 2023 fev 28]. Disponível em: <https://trends.google.com/home>
9. Jun S, Yoo HS, Choi S. Ten years of research change using Google Trends: from the perspective of big data utilizations and applications. *Technol Forecast Soc Change.* 2018;130:69-87. doi: <https://doi.org/10.1016/j.techfore.2017.11.009>

Recebido em 13/3/2023  
Aprovado em 22/3/2023