Google Trends Verification Study

Nathan Inkiriwang

The University of Sydney

Literature Review

In an age where internet searches are instantaneously recorded, a tremendous amount of current data is generated whose applications are practically limitless, and with Google's dominance in this sector over the past decade, there has been a substantial increase in the use of Google's publicly accessible Trends data for study in a wide range of disciplines (1). Numerous studies have utilised Google Trends (GT) to accomplish a variety of tasks in fields such as health sciences (1,2), social sciences (3,4), and the economy (5), with each study stressing the pros and cons that the usage of GT data has brought to their research.

Google Trends (GT) is a tool that Google made accessible to the public in 2006 providing users access to largely unfiltered *sample* of actual search requests made to Google. Standard GT studies would begin by obtaining search data from https://trends.google.com/, a website which allows researchers to select a search term or topic of interest and provides choices for filtering based on location, search type, time range, and search category. Once chosen, the website would create normalised search volume figures on a scale from 0 to 100 depending on a topic's fraction of total searches (6).

Initially, one of the most prevalent applications of GT data was in the health sector. In 2009, one of the early studies to utilise GT data, conducted by Google associates and the US Centres for Disease Control (CDC), revealed the value that GT data offered to the monitoring of health-related inquiries. The paper demonstrated how GT data surpassed the CDC's standard flu forecasts. (1 day latency for GT data versus 1-2 weeks for CDC) (7).

Since then, the use of GT data in health research has drastically risen (2), motivating some academics to create a comprehensive methodology framework that "provides a clear overview and specific direction for future researchers." (8) In light of the large growth of GT research, particularly in the field of health sciences, multiple recent papers have, however, highlighted the numerous issues and limitations that GT data presents. Google's normalisation and the lack of a exact scale and size for search volume, for example, make it extremely difficult to not only evaluate the significance of the changes that appear in search volume graphs, but also make it impossible for researchers to compare a large number of varied filtered search terms (On the GT website, you can only compare up to five terms, and you cannot compare between regions, search categories, etc.) (9).

Google's control over how GT data is outputted is the second limitation that has been raised in the past. The fact that Google has complete control over the data's origin, processing, and output poses a considerable hurdle to the study efforts of many scholars. Changes made to the website's functionality will make it exceedingly difficult to impossible for academics to repeat the findings of other researchers or even reproduce results over the course of an ongoing research period. Moreover, any algorithmic adjustment of Google's search algorithm may pose as a significant source of bias in GT study. For instance, Google's apparent usage of recommended searches to maximise ad revenue will increase the relative amount of specific search terms and constitute a significant source of bias and distortion in any data provided by GT (10)

The final and most significant issue with GT data is its extremely limited sampling method for each output. Google uses only 10-15% of the total search data it has for displaying results on its website (6). This is a serious challenge when attempting to conduct studies employing GT data to identify correlations or forecast regressions. The small data sample may significantly under or over-represent the search population. In addition, because we do not know the particular procedures Google used to obtain a 10-15% sample, it is hard for academics to be confident in the randomness of Google's sampling. One possible solution is to repeat the GT procedure multiple times in the hopes that Google will supply a different 10-15% sample each time. However, Google only updates the sampling for each GT search every 24 hours, and thus it took German researchers Vosen et al. 180 days to obtain 180 unique samples for their GT study (11). Recent research has, however, uncovered a way to use Google's API in order to collect several GT data samples with relative ease. Consequently, the question arises as to how significantly repeated sampling affects past and future studies (12,13)

Problem Statement

Due to insufficient data sampling, it is possible that the vast majority of Google trends studies conducted to date are inaccurate. Using Google Trends data as independent samples may have a substantial effect on the results of a study. This project attempts to replicate two studies, the Parable of Google Flu study by Lazer et al. (10)and the Methamphetamine study by (5) Gamma et al. in order to verify the validity of their results through multiple sampling. Specifically, the primary objective of this project is to replicate the studies using 130 samples of Google Trends data for the same search terms. We hypothesise that the study's published result will be verified from only a small selection of the samples, depending on the significance/confidence intervals of the original result, and thus the findings of this study will have implications for the credibility of varsity of previously published google trends research studies.

Methods

We begin by collecting 130 samples of Google Trends data for each search term utilised in the Parable of Google Flu study (10) and the Methamphetamine study (5) (Table 1).

Table 1 Search Terms

	Parable of Google Flu study	Methamphetamine study
Search terms utilised	"abdominal pain on my right side", "amoxicillin", "early signs of the flu", "fever", "influenza a", "pnumonia", "robitussin", "strep throat", "symptoms of bronchitis"	"meth"

In order to acquire these samples, we will collect one sample per day for 130 days. Within each request for a new sample, we will include the above-described search terms and the date period used in the two articles - 2004-2016 in the case of the Methamphetamine study (5) and 2004-2014 for the Parable of Google Flu Study (10). At the conclusion of the 130-day period, we would have acquired 130 unique files for each search keyword and region (One file for each sample we collect each day). We will then merge all files associated with the Methamphetamine study into one file and all files associated with the Google Flu parable study into a separate file, resulting in the creation of two data frames, one for the Parable research and one for the Methamphetamine study. This procedure was carried out using Python 3.9.7, and the detailed procedures are described in the appended supplementary file.

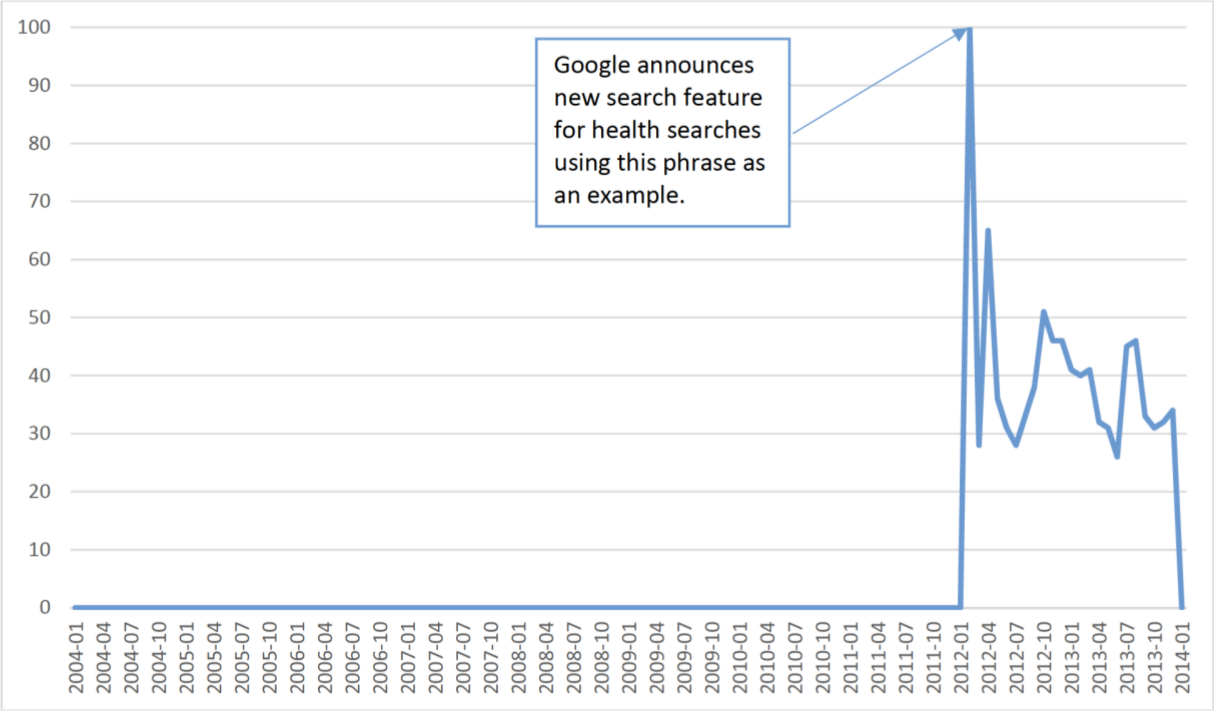
After integrating our 130 samples, we will need to combine our Methamphetamine data frame with the Methamphetamine study's crime data. This approach involved linking the two datasets based on matching regions and dates and was performed using R (RStudio 2022.07.1+554).

Once we have gathered our 130 samples, we will begin analysing the two studies, focusing on their published data, methodologies, and supplemental materials, in an effort to replicate their findings. Using their provided data sets, we will reproduce several selected plots (Fig S.10 and Fig. S15 from the Parable of Google Flu paper (10) and Fig. 1 and Fig. 3 from the Methamphetamine study (5)) in R (RStudio 2022.07.1+554). Following this, we will conduct the same procedure 130 times on the 130 downloaded samples of GT data to assess how multisampling affects the results. We will compare the individual results and values of the 130 GT samples to the results of the original study, as well as comparing the overall mean of our 130 samples to the single sample of the original study.

Results - Parable of Google Flu study

The first graph that will be replicated is Fig. S10 from the Parable of Google Flu Study (10) (Figure 1). The search volume for "abdominal pain on my right side" between 2004 and 2014 is illustrated in the graph below, which was obtained by the original researchers.

Figure 1 – Fig S.10 taken from Parable of Google Flu Study (10).

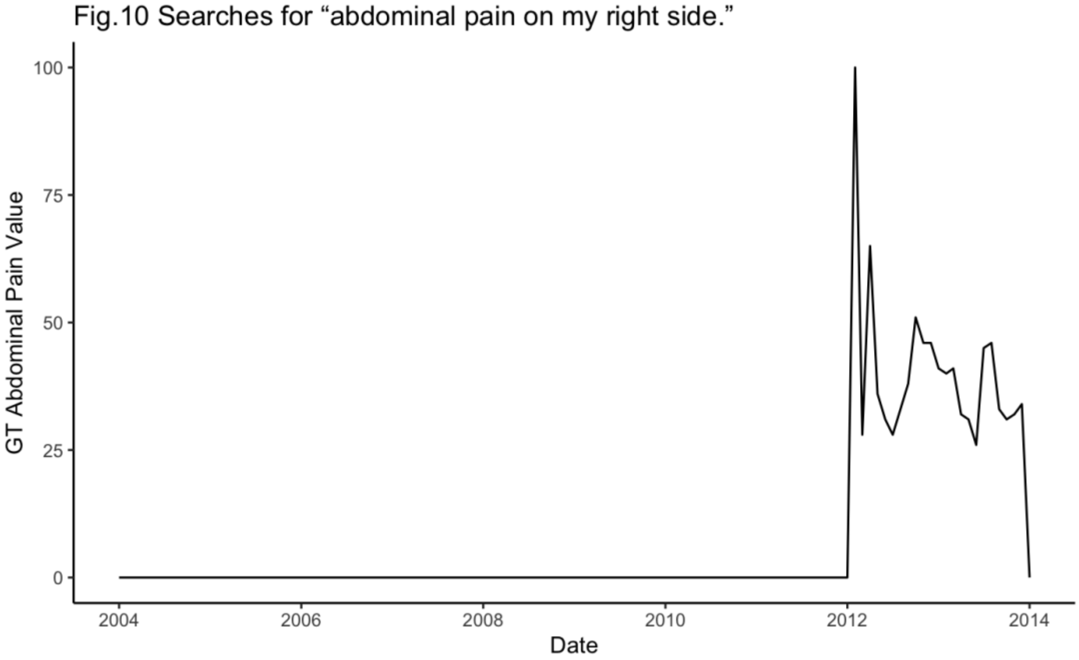


Google announces new search feature for health searches using this phrase as an example.

Source: Google Trends (http://www.google.com/trends/) (*39*). Downloaded data available in replication materials.

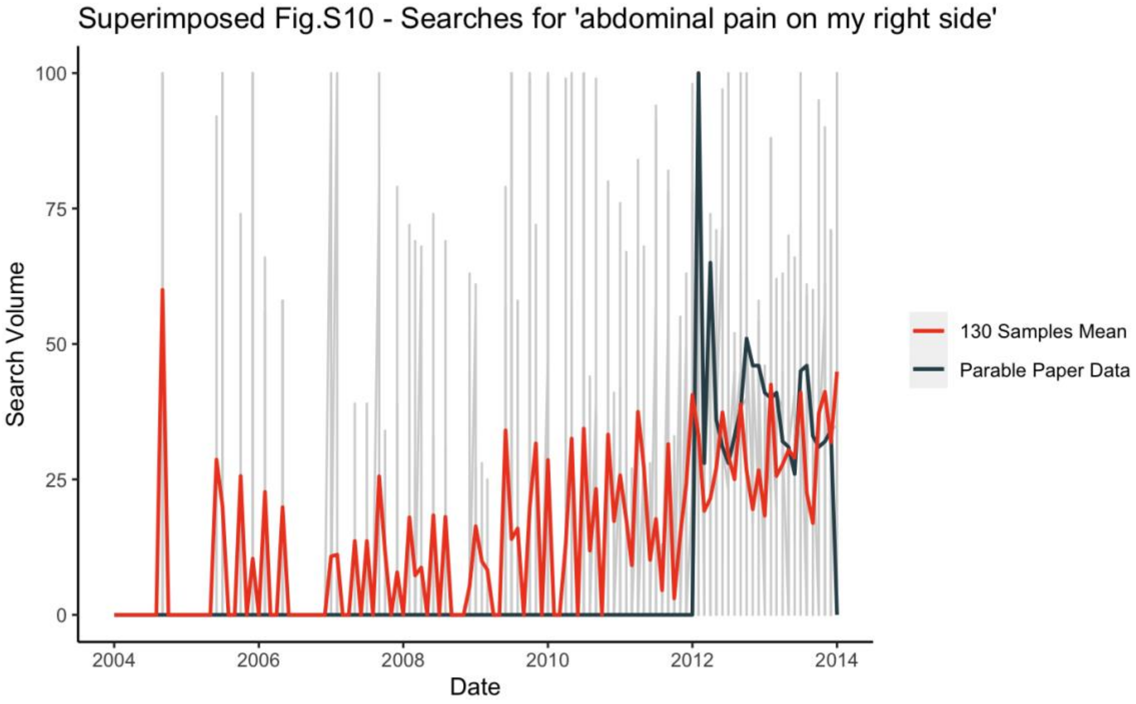
Using the original data supplied in the supplementary materials, we were able to duplicate the precise results of the Parable of Google Flu, as shown in the graph below. The source code for this R (RStudio 2022.07.1+554) procedure can be found in the appended extra file.

Figure 2 – Replicated Fig S.10 using original data.



Once we were able to replicate the results obtained by the Parable of Google Flu paper, we superimposed our obtained 130 samples of search term "abdominal pain on my right side" against the plot above. The resulting plot is shown below (Figure 3).

Figure 3 – Superimposed Fig S.10 using our obtained 130 Samples

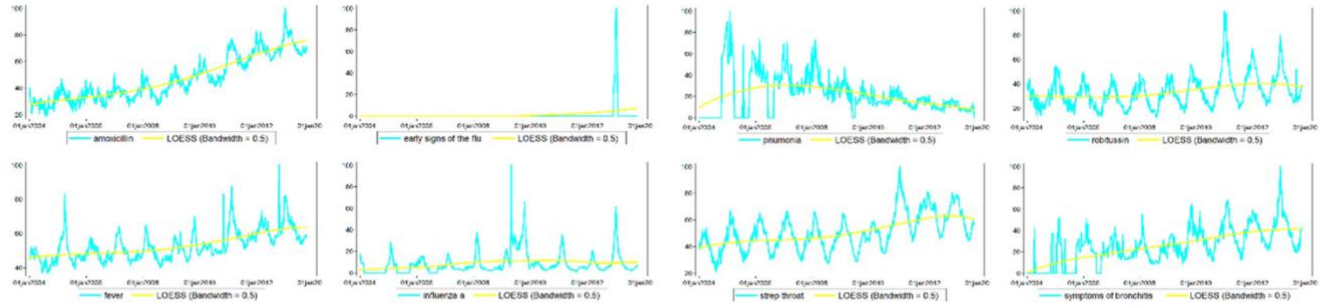


The plot reveals a considerable disparity between the findings produced from the parable paper data and the mean of our 130 samples. Using the mean of our multiple samples as opposed to the single sample utilised in the flu paper's data reveals a significant increase in overall variability. Particularly between 2004 and 2012, when the parable paper data appears to restrict extremely small search volumes to zero, our data could fluctuate far more. There may be a multitude of reasons that contributed to these considerable variations in outcomes, but one significant aspect may be the specificity of this search term. The extremely specific nature of the search query "abdominal pain on my right side" ensures that non-exact searches would not be counted. For instance, searches for "abdominal pain on the right side" would not be included because "my" is absent. As a result, the specificity of this search phrase may contribute to a very low total search volume, implying that considerably greater variation is anticipated across samples, hence supporting our initial hypothesis regarding the effect of multi-sampling.

Figure S15 from the Parable of Google Flu Study will be replicated as the second graph (10) (Figure 4). Similarly, this graph illustrates the distribution of search data values between 2004 and 2014. However, this figure does so for a variety of search terms as shown below.

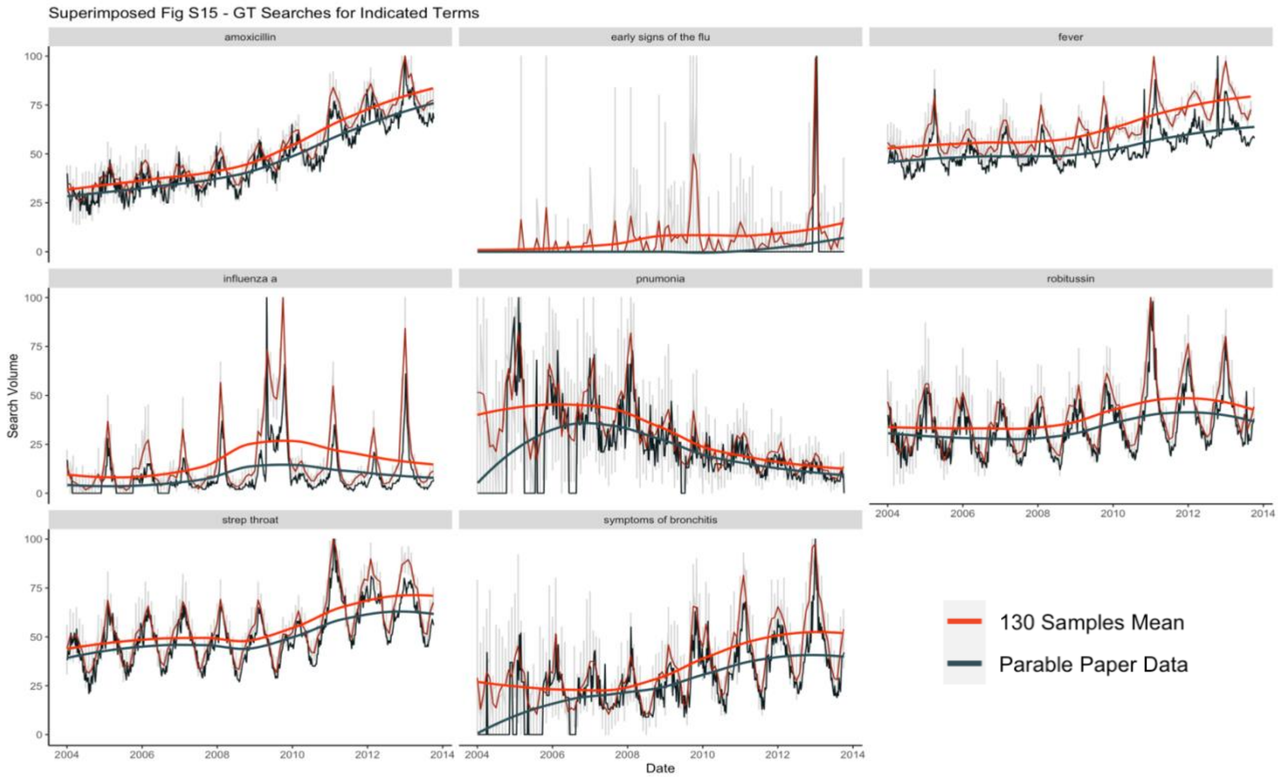
Figure 4 - Fig. S15 taken from the Parable of Google Flu Paper (10)

Fig. S15. Google Trend Searches for Terms Indicated in *PLOS One* Article (*14*). Data were acquired by typing search terms into Google Trends (http://www.google.com/trends/) (*39*) and downloading the associated data. The downloaded data are available in the SOM/SOM4/FigS15 folder of the replication materials.



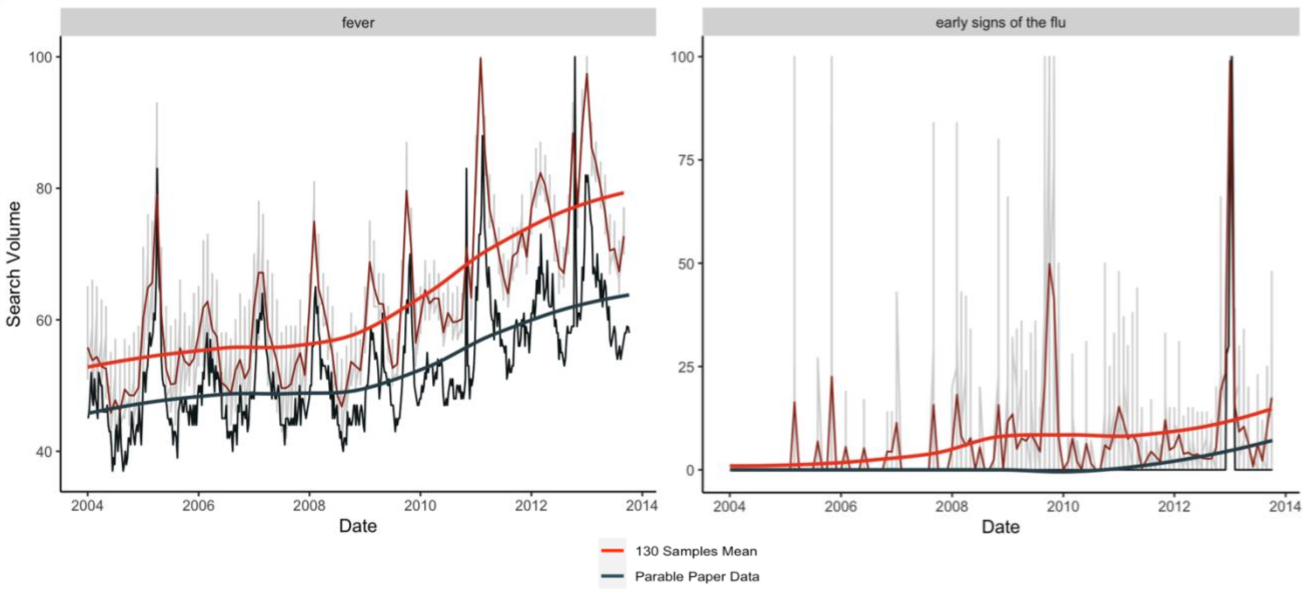
Using the original data provided in the supplementary materials, we were able to replicate the exact outcomes of the Parable of Google Flu and superimpose the distribution of our 130 samples, as depicted in the graph below (Figure 5). This R (RStudio 2022.07.1+554) procedure's source code can be found in the appended supplemental file.

Figure 5 - Superimposed Fig. S15 using our obtained 130 samples



Here, we can see that some search terms vary significantly more than others; consequently, we will select a few plots and investigate them individually. Below is the search term "fever" plot and "early signs of the flu" plot which will be the subject of our initial individual examination.

Figure 6 - Superimposed Fig. S15 for "Fever" and "early signs of the flu" search term/

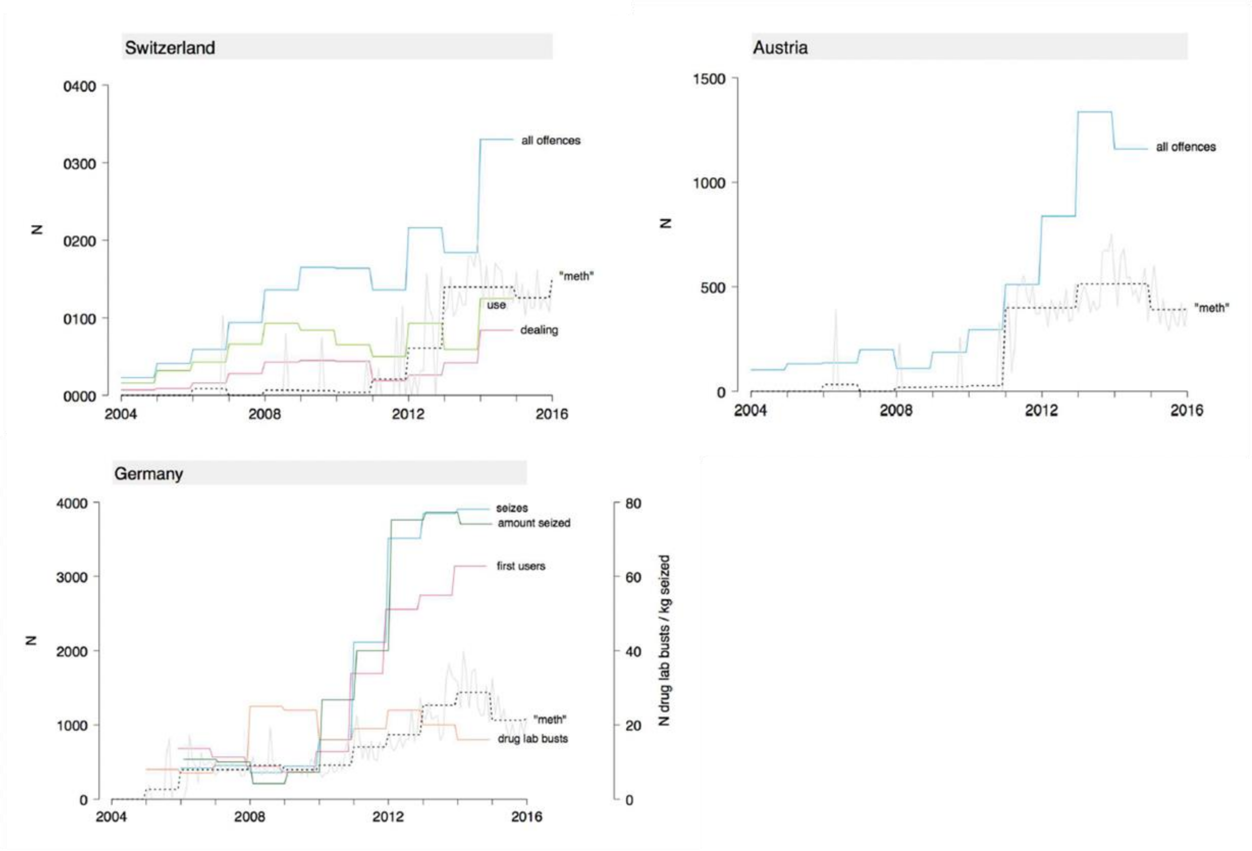


Despite the fact that the data from the parable paper appear to consistently underestimate our 130-sample mean for the "fever" search term, the distribution of search volume across years is remarkably similar. The same could not be said for the "early signs of the flu" plot, which reveals statistically significant differences between the parable of flu paper and our 130 samples. The extremely high discrepancies observed for the search term "early signs of the flu" but not for the search term "fever" support our previously mentioned factor of search term specificity. The search term "fever" is extremely inclusive, encompassing all queries containing the term "fever" (e.g., 'I have fever'). However, "early signs of the flu" is a very specific search term, resulting in a much lower overall search volume than "fever" and a much greater variation between samples. This demonstrates the critical requirement for multi-sampling, particularly when researching search terms with a relatively low search volume.

Results – Methamphetamine Study

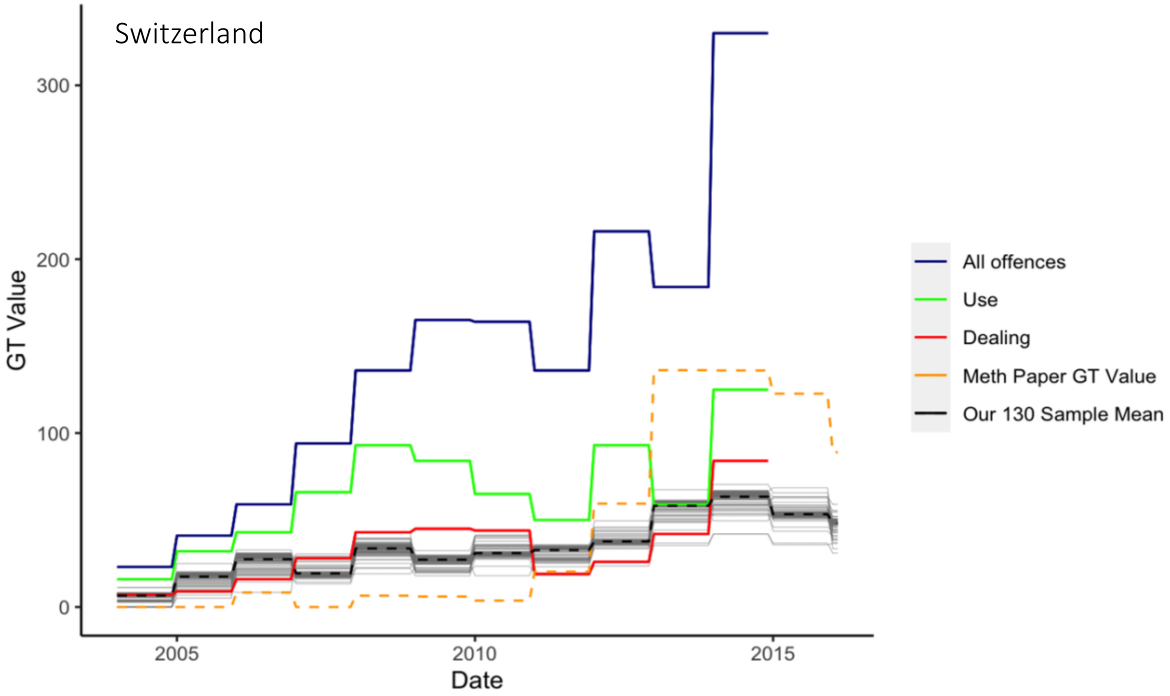
The first graph to be reproduced from the Methamphetamine Study (5) is Fig. 1 (Figure 7), which depicts the relationship between "meth" searches and meth-related crime by comparing the distribution of different meth-related crime variables derived from police data for Switzerland, Germany, and Austria to Google Trends search values for the term "meth" in those three countries. The graph obtained by the Methamphetamine researchers is shown below.

Figure 7 - Fig. 1 from Methamphetamine Study (5)



Using the original data supplied in the supplementary materials, we were able to reproduce the precise results of the Methamphetamine Study (5) and superimpose the distribution of our 130 samples. Following this, we will examine each of the three countries individually. The source code for this R (RStudio 2022.07.1+554) technique can be found in the appended supplementary file.

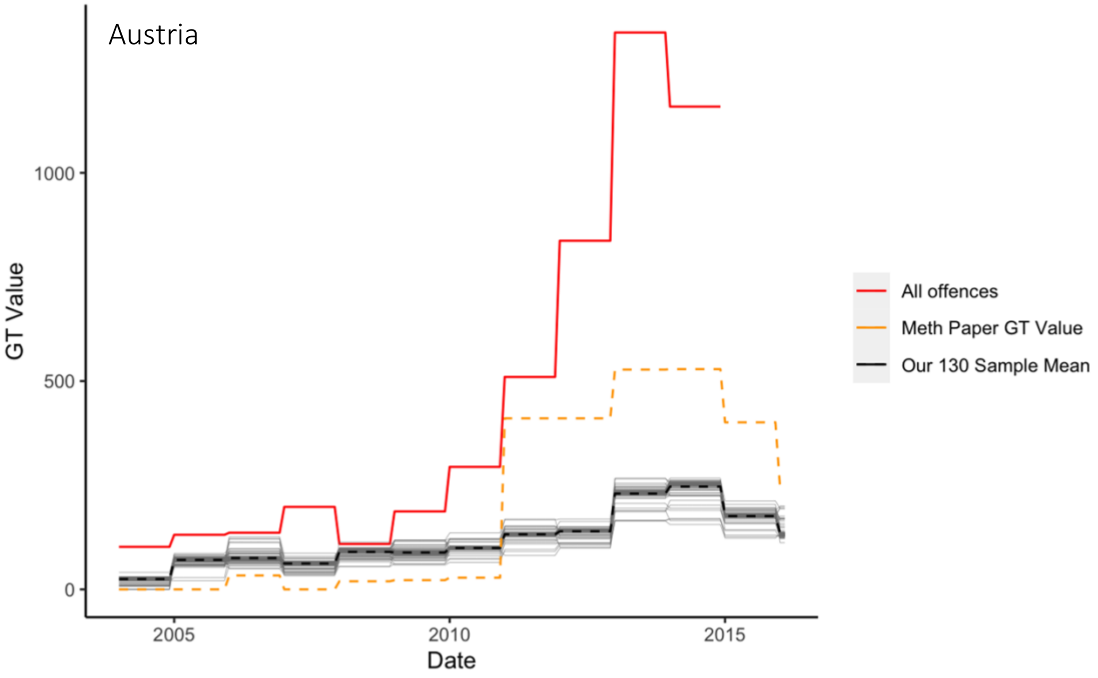
Figure 8 - Replicated and Superimposed Fig. 1 for Switzerland



The distribution of meth-related crime variables and meth searches in Switzerland is depicted in the graph above (Figure 8). Looking in particular at the yellow dashed line and the black dashed line, which emphasise the difference between our 130-sample mean and the Meth Paper data, we can see that the meth values we've obtained differ significantly.

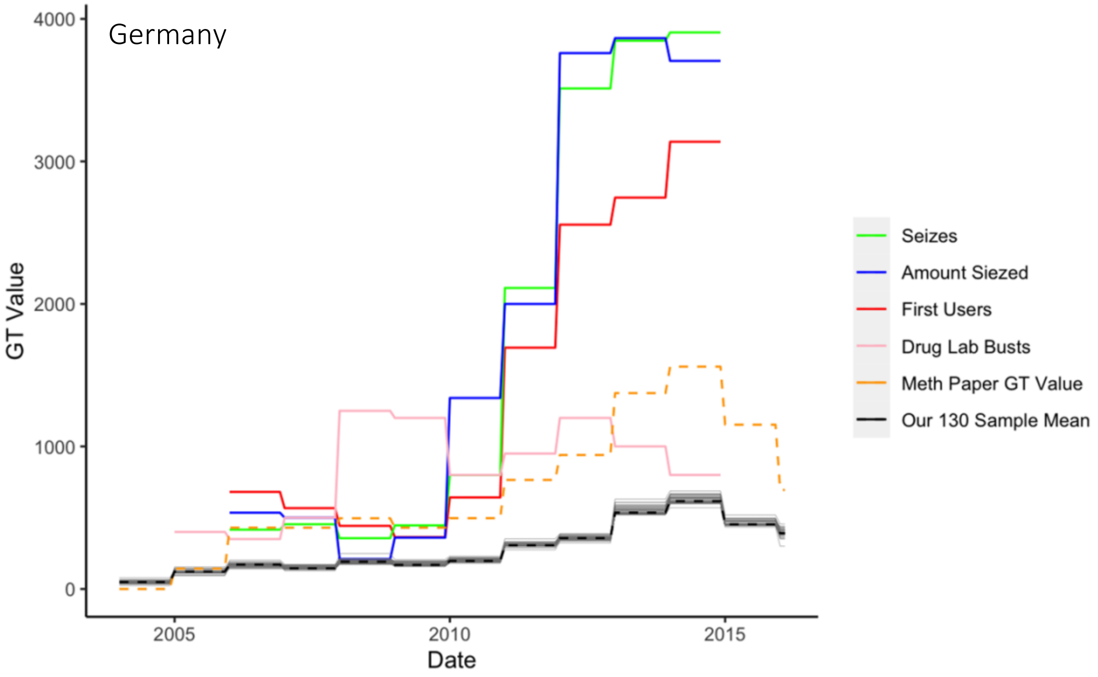
The Google Trend Values obtained in the single sample used by the Methamphetamine Study (5) appear to have underestimated the GT values of all of our samples from 2005 to 2012 and to have greatly overestimated our 130-sample mean for 2013 to 2016. Moreover, it is very interesting to note that none of our 130 individual samples (highlighted by the grey lines) come close to the values obtained by the Meth paper researchers, questioning the GT values obtained by the Methamphetamine Study researchers (5) and further accentuating the need for multi-sampling.

Figure 9 - Replicated and Superimposed Fig. 1 for Austria



Looking again at the dashed yellow and black lines from the plot above (Figure 9), a very similar pattern could be observed when examining the Austrian data, indicating the initial underestimation that evolves into a later overestimation.

Figure 10 - Replicated and Superimposed Fig. 1 for Germany



Lastly, while examining the plot for Germany, we can once again note the significant differences between the single sample of meth paper GT values and our 130-sample average. In contrast to the preceding scatter plots, the paper on meth appears to consistently overstate both the individual 130 samples and their mean. Compared to the plots of Switzerland and Austria, it appears that the variance between our 130 samples, shown by the grey lines, is much reduced in the Germany figure. Notate, however, that the scale of the Y-axis is significantly different from that of the Switzerland and Austria plots, which explains why the apparent variability is much lower.

The second graph to be reproduced from the Methamphetamine Study (5) is Fig. 3, which depicts the cross-correlation of Google search volume and criminal offences related to methamphetamine. The graph obtained by the Methamphetamine researchers is shown below.

Figure 11 - Fig. 3 from Methamphetamine Study (5)

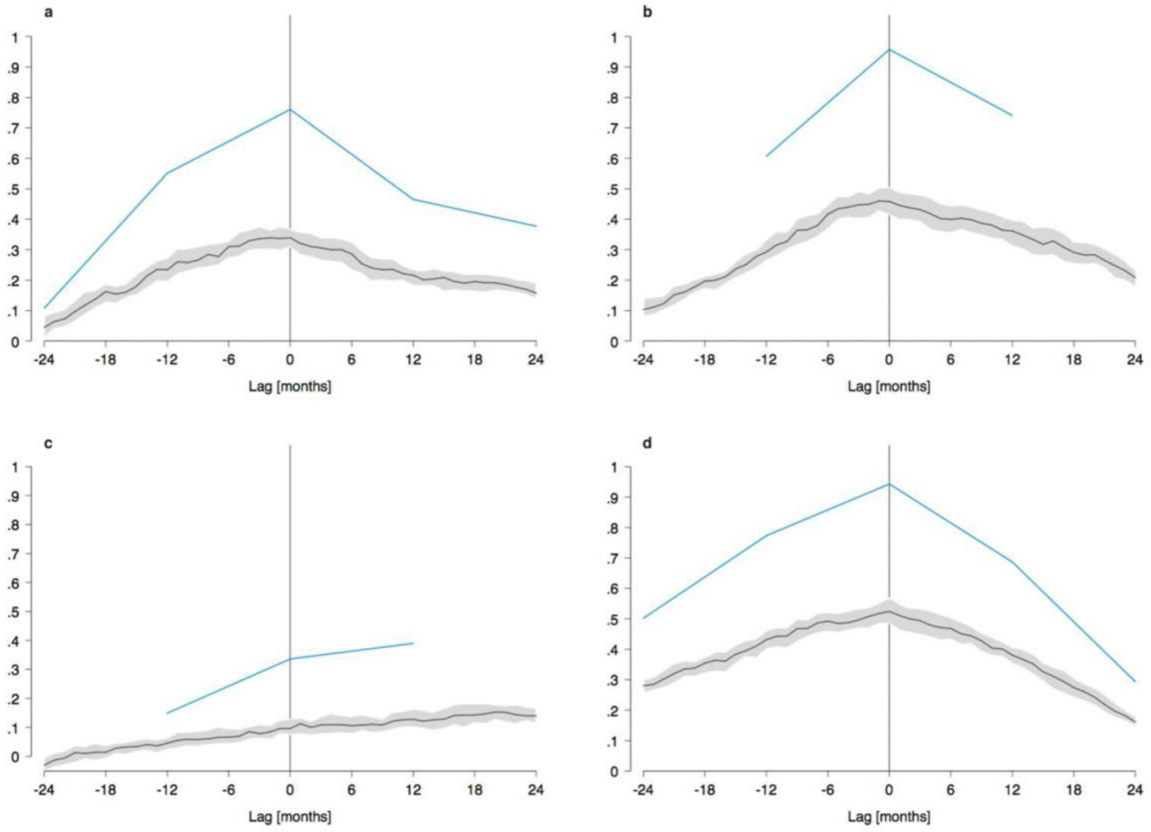
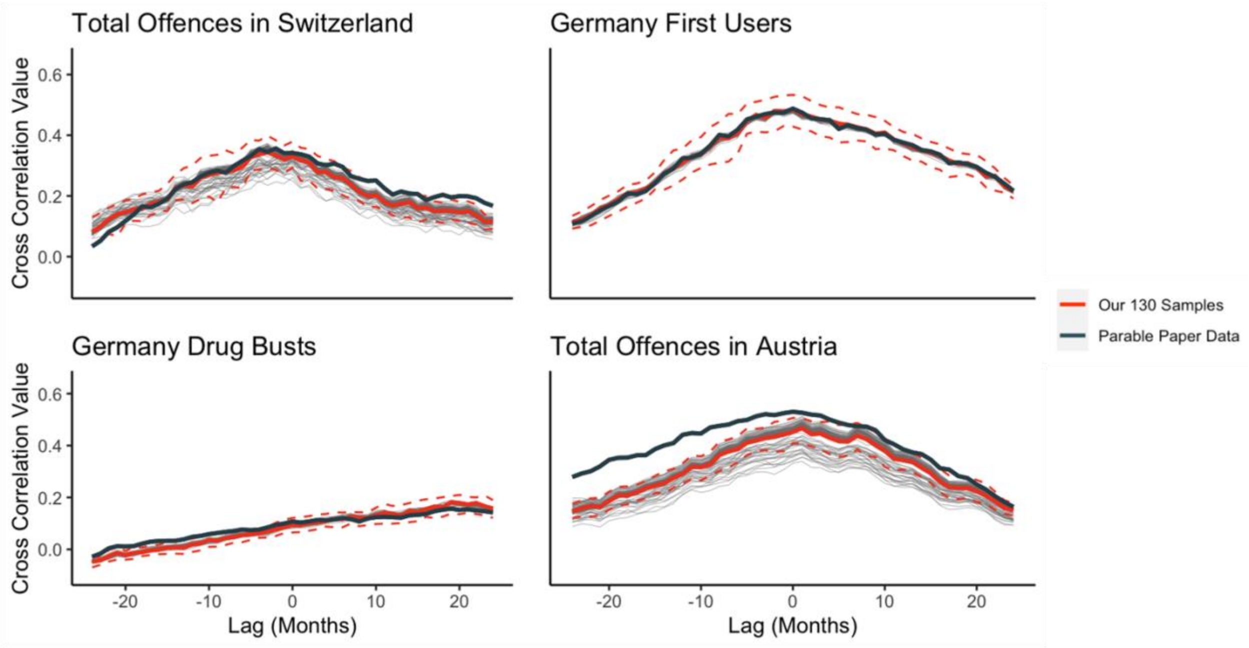


Fig 3. Cross-correlation of Google search volume and criminal offences related to methamphetamine. The blue curve shows cross-correlations based on yearly values. The gray line and shaded area represent the median and interquartile range of monthly correlations based on imputed data. a Total offences in Switzerland. b Number of first methamphetamine users in Germany. c Number of drug lab busts in Germany. d Total offences in Austria.

The figure above (Figure 11) illustrates the association between several criminal characteristics and the values of Google Trends that have shifted relative in time. It is also important to note that the Black line denotes the median, while the grey line illustrates the Interquartile Range.

Using the original data supplied in the supplementary materials, we were able to reproduce the precise results of the Parable of Google Flu and superimpose the distribution of our 130 samples, as depicted in the graph below. The source code for this R (RStudio 2022.07.1+554) technique can be found in the appended supplementary file.

Figure 12 - Superimposed Fig. 3 using our obtained 130 samples



Note that the Cross-correlation graphic utilises the Median and IQR, both of which are estimators with high reliability. Therefore, we should not anticipate substantial differences. Nonetheless, the above cross-correlation scatterplot reveals that some cross-correlations differ more than others. Specifically, when examining the total number of offences in Austria, the parable paper data exaggerated the presence of a link during earlier lags. Even the fact that, on average, the distribution of correlations between lags is quite similar across the majority of plots, we can still observe differences despite utilising the median.

Discussion

Our former sections revealed the inconsistencies between the four plots we selected from the Methamphetamine study (5) and the Parable of the Google Flu study (10). The majority of our superimposed plots corroborate our initial premise that just a small percentage of our 130 samples verified the article's published conclusions. In addition, we observed that for some plots (Figure 8, Figure 9), close to none of our 130 samples were able to replicate or confirm the conclusions of the researchers. This has led to the question of what could have brought about such substantial disparities.

The specificity of search terms is one likely factor correlating to the large variances indicated in the previous section. By comparing the “fever” and “early signs of the flu” plots (Figure 6), we have demonstrated that due to the small sample size, more specific searches tend to have greater deviation and variability in GT values. Consequently, this poses a problem for any researcher interested in conducting studies utilising precise terms. Particularly when examining GT values for specific search terms with low volumes, our findings indicate that researchers must employ multisampling. Due to the small number of searches, Google's sampling approach may not be sufficiently random to create an adequate sample that reflects the genuine population distribution.

In addition, the presence of unavoidable variation between each Google sample (more variation for specialised search phrases, less variation for broader search terms) may call into question the ethical usage of GT values. By only using one sample, as the two articles did, it is possible that researchers sampled multiple times to get the sample that best fits and supports their findings. Due to the inability of future researchers to replicate and verify the one sample of GT values that researchers have used (as Google always randomly selects samples and does not provide an option to set the randomisation seed), the accuracy of the findings is called into question. Using multi-sampling, however, ensures that researchers cannot just select the sample they believe best represents the population. Using the mean of a sufficiently large sample would ensure that the mean increasingly resembles the population mean, as indicated by the central limit theorem.

Multisampling always raises the question of what constitutes a "sufficiently large sample." To visualise how the variety of GT values varies as the number of samples varies, I developed an application that allows users to visualise the plots from the Parable of Google Flu study (5) and observe how varying the number of samples may impact the distribution of GT values. The application can be found here. In light of the variety between sample sizes, it is difficult to identify the exact minimum sample size for a sufficiently big sample. Nevertheless, based on our results and the app, we can assert with certainty that for more specific search terms, a larger sample size is required to accurately represent the genuine population mean. For larger search queries, researchers may be able to get away with fewer samples. On the basis of the law of large numbers, however, a sample size of $n \geq 30$ is recommended for GT data.

Insights

Even when using multi-sampling, there are some limits that must be considered. For starters, Google's complete control over all components of the GT website means that any modifications to sampling methods and algorithms could render all previous studies obsolete and unreproducible (10). Second, because internet searches cannot be devoid of biases and confounding variables, the assumption that the search volume for a given search phrase reflects what people are genuinely seeking is a glaring weakness of any Google Trends study (9). As a result, it is important to note that some researchers have turned to social media platforms such as Twitter, which provide significantly more context (due to the longer nature of tweets compared to Google searches) and enable researchers to perform natural language processing to better deduce the meanings of tweets (14).

Conclusion

In conclusion, we have proved the necessity of multisampling through our study by comparing the results achieved while using 130 samples vs a single sample. As anticipated, our use of multi-sampling drastically impacted both the Methamphetamine study and the Parable of Google Flu paper's outcomes, casting doubt on all previous Google Trends studies conducted without multi-sampling.

Future Work

Fortunately, we now have access to algorithms developed by researchers to quickly collect multisampling for Google Trends data, such as the one provided by Dr. Raubenheimer to obtain multiple samples using his published programme (12,13). As a result, future researchers have few justifications for not employing multisampling, and all future Google Trends studies should employ multisampling.

References

1. Arora VS, McKee M, Stuckler D. Google Trends: Opportunities and limitations in health and health policy research. Health Policy (New York). 2019 Mar 1;123(3):338–41.
2. Nuti S v., Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: A systematic review. Vol. 9, PLoS ONE. Public Library of Science; 2014.
3. Tran US, Andel R, Niederkrotenthaler T, Till B, Ajdacic-Gross V, Voracek M. Low validity of Google Trends for behavioral forecasting of national suicide rates. PLoS One. 2017 Aug 1;12(8).
4. Raubenheimer JE, Riordan BC, Merrill JE, Winter T, Ward RM, Scarf D, et al. Hey Google! will New Zealand vote to legalise cannabis? Using Google Trends data to predict the outcome of the 2020 New Zealand cannabis referendum. International Journal of Drug Policy. 2021 Apr 1;90.
5. Gamma A, Schleifer R, Weinmann W, Buadze A, Liebrenz M. Could Google Trends Be Used to Predict Methamphetamine-Related Crime? An Analysis of Search Volume Data in Switzerland, Germany, and Austria. PLoS One. 2016 Nov 30;11(11):e0166566.
6. Google. https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052.
7. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009 Feb 19;457(7232):1012–4.
8. Mavragani A, Ochoa G. Google trends in infodemiology and infoveillance: Methodology framework. Vol. 5, JMIR Public Health and Surveillance. JMIR Publications Inc.; 2019.
9. Ellery PJ, Vaughn W, Ellery J, Bott J, Ritchey K, Byers L. Understanding internet health search patterns: An early exploration into the usefulness of Google Trends. J Commun Healthc. 2008 Oct;1(4):441–56.
10. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science (1979). 2014 Mar 14;343(6176):1203–5.
11. Vosen S, Schmidt T. A monthly consumption indicator for Germany based on Internet search query data. Appl Econ Lett. 2012 May;19(7):683–7.
12. Raubenheimer JE. Google Trends Extraction Tool for Google Trends Extended for Health data. Software Impacts. 2021 May 1;8.
13. Raubenheimer JE. A Practical Algorithm for Extracting Multiple Data Samples From Google Trends Extended for Health. Am J Epidemiol. 2022 May 4;
14. Butler, D. (2013). When Google got flu wrong.