Harnessing Google Health Trends API Data for Epidemiologic Research: A Methodological Approach

Krista Neumann[1], Susan M. Mason[3], Kriszta Farkas[1], N Jeanie Santaularia[3,4], Jennifer Ahern[2], Corinne A. Riddell[1,2]

[1] University of California Berkeley, School of Public Health, Division of Epidemiology
[2] University of California Berkeley, School of Public Health, Division of Biostatistics
[3] University of Minnesota, School of Public Health, Division of Epidemiology and Community Health
[4] Minnesota Population Center, University of Minnesota

Introduction

Data from the **Google Health Trends Application Programming Interface (GHT-API)** can be useful for characterizing epidemiological patterns of exposure/disease. To access, researchers specify the search term(s), geographic region, and time period of interest, and the GHT-API returns an estimated scaled proportion of all Google searches.

However, there is little formal guidance about how to craft a GHT-API search strategy that will most accurately measure a construct of interest. Specific challenges include: 1) Data is suppressed when the number of searches is below a specific, undocumented threshold; and 2) Sampling variation due to the fact that GHT-API estimates proportions from a uniformly distributed random sample that is updated once a day.

Our objective is to describe best practices when using GHT-API to measure a construct of interest.

Motivating Case Study:

To examine trends in child abuse and neglect during the COVID-19 pandemic. Of concern is the possibility that pandemic-related challenges (e.g. school closures) may reduce the number of detected child abuse and neglect cases via traditional data sources, even if incidence of abuse and neglect increased. We thus investigated GHT-API as a real-time data source to capture state-level trends in child abuse and neglect.

Methods

To measure a construct of interest using GHT-API data, the following iterative search strategy approach is recommended:

1 — Identify how Google users search for construct of interest

Ascertain how the target audience describes the construct of interest in their Google searches and generate a comprehensive list of potential search phrases. A literature search and expert opinions may both be useful approaches for identifying relevant search phrases.

2 — Use incognito searches to improve sensitivity and specificity

Refine the comprehensive list from Step 1 so that phrases are broad enough to encompass the intended construct, but narrow enough to limit the number of false positives. Perform searches of each phrase using a Google Chrome incognito browser to identify both potential problematic phrases, and additional phrases for inclusion.

3 — Craft search term(s)

Combine individual (related) **search phrases** from step 2 into an overall **search term(s)** for querying the GHT-API using the Boolean OR operator "+". Combining related low-volume search phrases together can help improve low (or unstable) search volume and missingness.

4 — Determine feasible geographic and temporal scales

Test search term volume at the desired geographic-temporal resolution. To overcome a high degree of missing data, a larger geographic and/or time scale may be specified. The search term(s) may also need to be expanded (steps 1 - 3).

5 — Retrieve and average multiple samples to stabilize estimates

For a more stable estimate of the true underlying search proportion, perform multiple queries (over multiple days) to generate different random samples and average over resulting search volume estimates. This can also help to address data suppression not remedied by Steps 3 and 4 since GHT-API results which are close to the suppression threshold in one sample may fall above the threshold in another sample.

6 — *(optional)* Normalize to account for changes in total search volume

Normalization may help reduce changes in the proportional search volume that are due to, or an artifact of, changes in the total searches (Stephens-Davidowitz 2013). The aim is to select a normalization term for which searches are hypothesized to remain constant over the study period; however, in practice, there is no way to test this hypothesis.

Case Study Application

'[perpetrator] + [verb] + [victim]'
(e.g., 'mom hit me')

Step 1: We used a literature search and validated survey instruments to generate a series of abuse phrases

Step 2: We removed terms with irrelevant results and added additional perpetrators

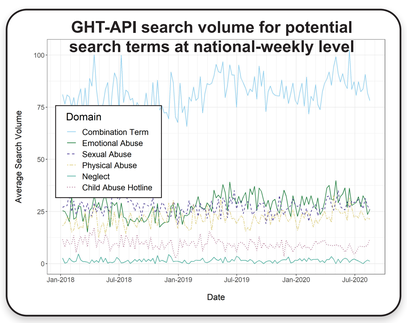
Physical abuse: 'mom hit me + dad kicks me + parent slapped me + ... '

Emotional abuse: 'mom hates me + dad threatens me + parent scares me + ... '

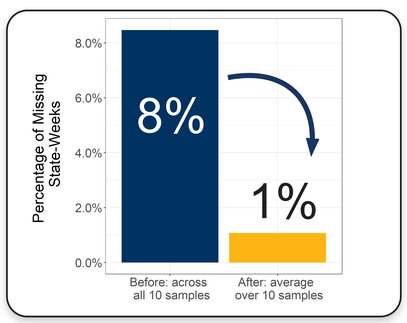
Sexual abuse: 'mom molested me + dad rapes me + parent has sex with me + ... '

Neglect: 'mom drinks a lot + dad uses drugs + parent left me alone + ... '

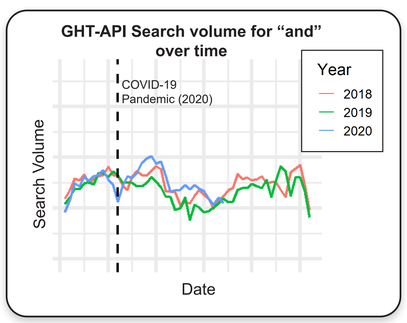
Step 3: We combined multiple search phrases into 4 GHT-API search terms by concatenating all abuse phrases within each abuse subtype



Step 4: Due to low search volumes at the national-weekly level, we concatenated all 3,484 separate abuse phrases into a single combined search term for analysis



Step 5: Averaging 10 separate state-week samples improved estimate stability and reduced missingness (8% to 1%).



Step 6: We normalized by "and" in a sensitivity analysis, but were unable to determine whether it helped overcome changes in total search volume

Conclusions

GHT-API data has the potential to provide early signals for phenomena in which reporting data is delayed, and for health outcomes that are difficult to measure by traditional mechanisms.

We provided a methodological framework and best practices for utilizing the GHT-API to study epidemiologically-relevant constructs.

Contact

For questions: krista_neumann@berkeley.edu

For details and code related to the case study please visit:
https://github.com/corinne-riddell/SIP-and-abuse