COVID-19 Search Trends symptoms dataset

Updated Feb 24, 2021

Terms of use

To download or use the data, you must agree to the Google Terms of Service.

Summary

This aggregated, anonymized dataset shows trends in search patterns for symptoms and is intended to help researchers to better understand the impact of COVID-19.

Public health experts indicated that trends in search patterns might be helpful in broadly understanding how COVID-19 impacts communities and even in detecting outbreaks earlier. You shouldn't assume that the data is a recording of real-world clinical events, or use this data for medical diagnostic, prognostic, or treatment purposes.

To visualize the data, try exploring these interactive charts and map of symptom search trends.

About the data

This data reflects the volume of Google searches for a broad set of health symptoms, signs, and conditions. *To keep things simple in this documentation, we will refer to all of these collectively as symptoms*. The data covers hundreds of symptoms such as *fever*, *difficulty breathing*, and *stress*—based on the following:

- a symptom's prevalence in Google's searches
- data quality and privacy considerations

For each day, we count the searches mapped to each of these symptoms and organize the data by geographic region. The resulting dataset is a daily and/or weekly time series for each region showing the relative frequency of searches for each symptom.

A single search query can be mapped to more than one symptom. For example, we map a search for "acid reflux and coughing up mucus" to three symptoms: *Cough*, *Gastroesophageal reflux disease*, and *Heartburn*.

The dataset covers the recent period and we'll gradually expand its range as part of regular updates. Each update will bring the coverage to within three days of the day of the update.

Although we are releasing the dataset in English, we count searches in other languages. In each supported country, we include the languages needed to cover the majority of symptom search queries. For example, in the United States we support Spanish and English.

The data represents a sample of our users and might not represent the exact behavior of a wider population.

Preserving privacy

For this dataset, we use differential privacy, which adds artificial noise to our datasets while enabling high quality results without identifying any individual person.

To further protect people's privacy, we ensure that no personal information or individual search queries are included in the dataset, and we don't link any search-based health inferences to an individual user. More information about the privacy methods used to generate the dataset can be found in this report.

How we process the data

We'd like to report symptoms for each day, but sometimes we can't do this. When the daily volume of the data for a given region does not meet quality or privacy thresholds, we do the following:

1. Try to provide a given symptom at the weekly resolution.
2. If we cannot meet our quality or privacy thresholds at the weekly resolution, we don't provide the data for the symptom in that region.

As a result, in a given region, some symptoms are available at a daily resolution while others are only available at weekly resolution. To make it easier to compare a wider range of symptoms within the same region, whenever daily data is available we also produce an aggregate weekly value computed from the individual daily (Monday to Sunday) values. We use this reaggregation approach for privacy reasons, as we cannot directly compute both daily and weekly versions of the same data. We refer to these values as weekly-from-daily.

Due to the reaggregation, the weekly-from-daily data has slightly more noise than the weekly data computed directly. The absolute magnitude of errors in the weekly-from-daily data time series is limited: under 15% for all the symptoms (aggregated over all weeks and locations), and under 10% for most symptoms. The errors are symmetrically distributed, suggesting that the weekly-from-daily values provide an unbiased estimate of the true weekly average.

If a symptom-region pair is available in both the daily and weekly time series, then the weekly estimates are aggregated from the daily values. Otherwise, they're computed directly.

With the addition of the weekly-from-daily values, in a given region, symptoms might appear in both the daily and weekly time series, only in the latter, or neither.

The data shows the *relative popularity* of symptoms in searches within a geographical region.

To normalize and scale the daily and the weekly time series (processed separately), we do the following for each region:

1. First, we count the number of searches for each symptom in that region for that day/week.
2. Next, we divide this count by the total number of Search users in the region for that day/week to calculate relative popularity (which can be interpreted as the probability that a user in this region will search for the given symptom on that day/week). We refer to this ratio as the *normalized popularity* of a symptom.
3. We then find the maximum value of the *normalized popularity* across the entire published time range for that region, over all symptoms using the chosen time resolution (day/week). We scale this maximum value to 100. All the other values are mapped to proportionally smaller values (linear scaling) in the range 0-100.
4. Finally, we store the scaling factor and use it to scale values (for the same region and time resolution) in subsequent releases. In future updates, when a symptom popularity exceeds the previously-observed maximum value (found in step 3), the new scaled values may be larger than 100.

For each region and time resolution, we scale all the normalized popularities using the same scaling factor. In a single region, you can compare the relative popularity of two (or more) symptoms (at the same time resolution) over any time interval. You can also compare a weekly-from-daily value with another weekly value because they share the same scaling factor. However, you shouldn't compare the values of symptom popularity across regions or time resolutions — the region- and time-resolution-specific scalings make these comparisons meaningless.

Note: We adopted a new scaling factor for the US weekly data across all symptoms starting on Dec 15, 2020. While the numbers for normalized search volume changed on this date, the normalized search volumes retain their interpretation relative to each other.

Data fields

The dataset includes the following fields:

- **country_region**: The name of the country in English. For example, *United States*.
- **country_region_code**: The ISO 3166-1 code for the country. For example, *US*.
- **sub_region_1**: The name of a region in the country. For example, *California*.
- **sub_region_1_code**: A country-specific ISO 3166-2 code for the region. For example, *US-CA*.
- **sub_region_2**: The name of a subdivision of the region above. For example, *Santa Clara County*.
- **sub_region_2_code**: For the US - the FIPS code for a US county (or equivalent). For example, *06085*.
- **place_id:** A textual identifier that uniquely identifies a place in the Google Places database and on Google Maps (details). For example, *ChIJd_Y0eVIvkIARuQyDN0F1LBA*.
- **date:** The day on which the searches took place. For weekly data, this is the first day of the 7-day weekly interval starting on Monday. For example, in the weekly data the row labeled *2020-07-13* represents the search activity for the week of July 13 to July 19, 2020, inclusive. Calendar days start and end at midnight, Pacific Standard Time.
- **<Symptom name>:** Repeated for each symptom. Reflects the normalized search volume for this symptom, for the specified date and region for example, *87.02*. The field may be empty when data

is not available.

Availability

To start working with the dataset (or just explore), you can do the following:

- Explore or download the data using our interactive charts.
- Run queries in Google Cloud's COVID-19 Public Dataset Program.
- Analyze the data alongside other covariates in the COVID-19 Open-Data repository.

We'll continue to update this dataset while public health experts find it useful in their work to stop the spread of COVID-19. We will also take into account feedback from public health researchers, civil society groups, and the communities at large.

Attribution

If you publish results based on this dataset, please cite as:

> Google LLC "Google COVID-19 Search Trends symptoms dataset".
> http://goo.gle/covid19symptomdataset, Accessed: <date>.

Feedback

We would love your feedback on the dataset and documentation, or any unexpected results.
Please email your feedback to covid-19-search-trends-feedback@google.com.

Dataset changes

Feb 24, 2021 - Added Place IDs to the dataset
Dec 15, 2020 - New regions, aggregate-weekly data derived from daily data, rescaled weekly data for United States, and CSV downloads from interactive charts
Sep 18, 2020 - New interactive charts and map of the dataset
Sep 02, 2020 - Initial release