COVID-19 Vaccination Search Insights

Updated August 30, 2022

Terms of use

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

Summary

This aggregated, anonymized data shows trends in search patterns related to COVID-19 vaccination. We're making this data available because we heard from public health officials that trends in search patterns could help to design, target, and evaluate public education campaigns.

About this data

These trends reflect the *relative interest* of Google searches related to COVID-19 vaccination. We split searches, by information need, across 3 categories:

1. **COVID-19 vaccination**. All searches related to COVID-19 vaccinations, indicating the overall search interest in the topic. For example, "when can i get the covid vaccine" or "covid booster shot". This parent category includes searches from the following 2 subcategories.
2. **Vaccination intent**. Searches related to eligibility, availability, and accessibility of vaccines. For example, "covid vaccine near me" or "safeway covid vaccine".
3. **Safety and side effects**. Searches related to the safety and side effects of the vaccines. For example, "is the covid vaccine safe" or "pfizer vaccine side effects".

We selected these categories based on the input from public health experts, as well as taking into consideration:

- data quality—for example, clear user intent
- privacy—for example, significant search volumes

A search classified in a subcategory, always also counts towards the parent *COVID-19 vaccination* category, however some COVID-19 vaccination searches may be classified in the parent category but neither subcategory (an example would be queries about COVID-19 vaccine brands).

For each of the 3 categories we also publish the top and rising weekly-trending search topics:

- **Top searches**. The 20 most-searched-for topics, ranked by the number of people searching for the term.

- **Rising searches**. The next 20 most-searched-for topics that have at least a 50% increase over the previous week. We rank the searches by the number of people searching for the term.

The data covers the period starting Jan 2021 to the present. We'll offer weekly updates covering the most recent week (Mon–Sun). Each update will be available a few days after the week ends—to allow time for data processing and validation.

This data represents Google Search users and might not represent the exact behavior of a wider population. However, we expect that any systematic regional biases will remain stable over the period covered by the dataset.

How we process the data

The data shows the *relative interest* in each of the search categories within a geographical region. To generate, normalize, and scale the weekly time series we do the following for each region—let's consider an example region A:

1. First, we count the queries classified in each of the categories in region A for that week. To determine the region, we estimate the location where the query was made. When counting the queries, a given anonymous search user can contribute at most once to each category per day, and to at most 3 different categories per day.
2. Next, we divide this count by the total volume of queries (on any topic, not just those related to COVID-19 vaccination) in region A for that week to calculate relative interest. We call this proportion the *normalized interest* of a category. This is a relatively small number, which reflects the fraction of all search queries in that region that are related to the topic of COVID-19 vaccination or one of its subcategories.

   (Initial release only) We establish a *fixed scaling factor* by finding the maximum weekly value of the *normalized interest* for the general *COVID-19 vaccination category,* at the US national level (which occurred on the week starting at March 8, 2021). We scale this maximum value to 100 by multiplying it by a number which we set as the *fixed scaling factor*. We store the fixed scaling factor, and in subsequent updates we use it to scale values in all regions in every country.

3. Finally, using our fixed scaling factor, we linearly scale all the other normalized interest values, across regions, categories, and time. These values can be lower or higher than 100 (but not less than 0). We call these values *scaled normalized interest.*

For the weekly trending searches, we do the following for each region:

1. For each category, our differentially private algorithm counts how many anonymized users made each contributing query during the week. An anonymous user can contribute to 1 query each day.
2. Next, the process differs between query clusters groups and synonymous queries groups, depending on the country:
    a. For **query clusters**, we group queries by topic using k-means clustering. Queries that are close in meaning and intent will be in the same group—while other groups cover different intents and information needs. For each group, we total the user count (from the previous step) and associate the group with the most representative (*frequent*) query.
    For example, a group which contains [("omicron vaccine", count=400), ('vaccine effectiveness omicron", count=200), ("omicron vaccine pfizer", count=100)], gives ("omicron vaccine", count=700).
    Because we repeat this step every week, a group for a given representative query might contain different queries from week to week—although with similar meaning and intent.
    b. For **synonymous-query groups**, we lemmatize the search queries and combine them with synonymous queries. For each group, we total the user count (from the previous step) and associate the synonyms with the most representative (*frequent)* query.
    For example, a group which contains [("covid booster", count=200), ("booster shot", count=100), ("covid booster shot", count=300)] gives ("covid booster shot", count=600).
3. For *Top searches*, we rank the groups by count and report the most popular ones (up to 20).
4. For *Rising searches*, we exclude the *Top searches* and any group that didn't increase by 50% or more over the previous week, re-rank, and report the most popular ones (up to 20).
5. For reporting, we scale and normalize the group counts the same way as the counts of the 3 main categories, using the same fixed scaling factor. To show the breadth of a topic (only in query clusters), we also report a sample of popular search queries.

Because all *scaled normalized interest* values (across the 3 main categories and the trending searches) share the same scaling factor globally, you can do the following:

- Compare interest between categories or trending searches across all regions (even across countries) over any time interval.
- Calculate the fraction of COVID-19 vaccination queries that focus on the topic of vaccination intent or safety and side effects. To do this for a region, divide the *scaled*

normalized interest of the *Vaccination intent* or *Safety and side effects* categories by that of the *COVID-19 vaccination* category.

- Compare the ranking of the trending searches across locations and across time. Remember, a difference in ranking doesn't necessarily mean a change in interest (or absolute counts) for that search.

Sometimes it's not possible to report trends for every region. When the weekly volume of data for a given region doesn't meet quality or privacy thresholds, we cannot provide data for some or all categories in that region. In such cases, the data for that region will still be counted in its parent region. For example, data for all the counties in the US state of Nebraska will be counted as part of Nebraska's state trends. Because we omit the data for regions where the search volume doesn't meet our quality or privacy thresholds, we compute the data for each region directly from all the queries associated with that region, instead of using the aggregate data of its subregions.

How we classify search queries

Classifying web-search queries is challenging. Each query is a few words that can be illusory and ambiguous. So, we look at other signals beyond the query—especially the words and phrases found in the search results.

We use supervised machine learning to find the search queries that match the 3 categories. For each of the 3 categories, we trained a neural-network model with a single hidden layer. Each model has 60,000 input nodes, corresponding to words and phrases extracted from the query (and its associated search results) using information-gain criterion. We also added features to the model using entities associated with the query and/or its search results (similar to this Google Cloud entity analysis).

Using both the query and its search results allows our classifiers to automatically adapt their classifications to the current context and intent of the query. However, because the results for a query can change, a query such as "pfizer vaccine" might be classified as *COVID-19 vaccination* and a few months later also as *Safety and side effects*. To learn more about Google Search results, visit How Search works.

Table 1 shows the top features we used for each category. Some of the features are common across the COVID-19 Vaccination parent category and the subcategories.

Table 1. Top features used for each category

Category	Top features
COVID-19 vaccination (CA)	*vaccination, vaccines, vaccinations, vaccine, covid vaccine, vaccinated, covid, covid 19, immunization, 19 vaccine, covid vaccination, pfizer, vaccins, vaccin, vacciner, vaccinate, shots, booster shot, booster dose, vaccin contre*

COVID-19 vaccination (IE, UK, US, AU)	*covid vaccine, vaccines, vaccination, vaccinations, 19 vaccine, vaccine, vaccinated, covid 19, covid, coronavirus vaccine, immunization, coronavirus, covid vaccines, vaccine appointment, pfizer, health, pharmacy, second dose, cdc, doses*
Vaccination intent (AU)	*clinic, clinics, vaccination clinic, appointment, book, vaccine clinic, booking, vaccination centre, vaccination clinics, vaccine appointment, appointments, registration, register, vaccine registration, medical centre, vaccine clinics, vaccination centres, vaccination hub, astrazeneca vaccine*
Vaccination intent (CA)	*clinic, clinics, vaccine clinic, appointment, vaccination clinic, vaccine appointment, vaccination centre, vaccination centres, vaccination clinics, pharmacy, appointments, booking, vaccine centre, rendez vous, clinique, walk in, vaccine registration, walk in, prendre rendez*
Vaccination intent (IE, UK)	*appointment, appointments, book, booking, vaccination centre, clinic, vaccination centres, vaccine appointment, clinics, coronavirus covid, walk in, coronavirus vaccination, covid vaccination, vaccination clinic, vaccine clinic, centres, vaccination appointment, vaccine centre, centre, book covid, pfizer, astrazeneca*
Vaccination intent (US)	*pharmacy, pfizer, vaccine appointment, appointment, pharmacies, moderna, dose, appointments, pfizer vaccine, cvs, walgreens, second dose, vaccine appointments, cvs pharmacy, doses, shot, cvs covid, walgreens pharmacy, vaccine eligibility, moderna vaccine*
Safety and side effects (CA)	*side effects, vaccine side, side effect, symptoms, adverse effect, allergic reaction, pain, effet secondaire, effets indesirables, adverse reactions, effets secondaires, effet indesirable, reactions, chest pain, inflammation, nausea, fever, heart inflammation*
Safety and side effects (IE, UK, US, AU)	*side effects, side effect, symptoms, fever, second dose, allergic reaction, moderna injection, pfizer, reactions, reaction, pfizer vaccine, pain, health, shot, pharmacy, allergic reactions, adverse effects, adverse reactions*

Training our classifiers

We trained each country's models in a supervised manner using a sample of search queries made there during 2021— the period typically being a few months. We labeled the training data using a set of simple rules.

To develop the rules, we started by sampling a set of top queries that are associated with web pages about Covid-19 vaccines, Covid-19, or any vaccines. We manually marked each sample query as positive or negative against the three categories. For each category, we created rules from terms, phrases, and entities associated with the positive queries and rarely associated with

the negative queries. For example, for the *COVID-19 Vaccination* category we require "vaccine" and "covid" to be among the top most relevant terms. Finally we used these rules to automatically label the rest of the training data.

Evaluating our classifiers

To evaluate our classifiers' capacity to detect queries about COVID-19 vaccinations, we relied on Google's *search quality raters* who have deep experience with how health-related information needs are reflected in search queries. These raters were unknown to and independent of the developers of the classifiers. The raters were not aware of this project and did not know the purpose of their task.

Because only a small minority of Google Searches are for COVID-19 vaccination topics, we needed to create a sample set of queries for evaluation. We used Google Knowledge Graph entities to find queries which included high confidence positives, potential positives, and close negatives. For example, for the classifier used for *COVID-19 vaccination* category, we sampled top and random queries associated with the entity "Covid-19 Vaccination" (high precision), as well as queries that are only associated with the entity "Covid-19" or with "Vaccination" (high recall).

Table 2 shows the distribution of query ratings for each category. A neutral rating means either multiple raters entered a neutral rating for the query or there was no consensus. Queries that are rated as neutral are excluded from the classifier evaluation.

Table 2. Distribution of query ratings for the categories

Country	Category	Positives	Negatives	Neutral	Krippendorff's alpha
AU	Covid-19 vaccination	753	378	64	0.907
	Vaccination intent	286	804	105	0.819
	Safety and side effects	286	815	94	0.883
CA	Covid-19 vaccination	1642	703	267	0.730
	Vaccination intent	382	1968	262	0.692
	Safety and side effects	618	1710	284	0.849
IE	Covid-19 vaccination	528	411	193	0.860
	Vaccination intent	200	757	175	0.641
	Safety and side effects	170	838	124	0.810
UK	Covid-19 vaccination	1149	672	156	0.863
	Vaccination intent	264	1523	190	0.727
	Safety and side effects	498	1256	223	0.846

US	Covid-19 vaccination	1973	1122	337	0.844
	Vaccination intent	419	2724	289	0.713
	Safety and side effects	826	2183	423	0.811

The three raters independently judged the relevance of each search query in our sample to each of the three categories. The inter-rater agreement (measured by Krippendorff's alpha in table 2) indicates high agreement.

Table 3 shows that the classifiers achieved high precision as well as high recall when identifying queries related to each of the categories.

Table 3. Precision and recall scores for the classifiers

Country	Category	Precision	Recall
AU (English)	Covid-19 vaccination	0.98	0.91
	Vaccination intent	0.85	0.84
	Safety and side effects	0.97	0.85
CA (English & French)	Covid-19 vaccination	0.96	0.95
	Vaccination intent	0.89	0.87
	Safety and side effects	0.94	0.80
IE (English)	Covid-19 vaccination	0.94	0.91
	Vaccination intent	0.92	0.81
	Safety and side effects	0.94	0.94
UK (English)	Covid-19 vaccination	0.98	0.96
	Vaccination intent	0.84	0.80
	Safety and side effects	0.87	0.90
US (English)	Covid-19 vaccination	0.96	0.94
	Vaccination intent	0.83	0.81
	Safety and side effects	0.87	0.89

Preserving privacy and quality

To preserve user privacy, we use the state-of-the-art differential privacy approach, which adds artificial noise to our data and allows us to produce high quality data without identifying any individual person.
To further protect users' privacy, we ensure that no personal information is included in the data, and we don't link any related search-based inferences to an individual user.

To ensure accuracy after adding noise, we estimate the magnitude of change due to the noise. For the 3 main categories, we retain all the values that (after the addition of noise) have 80% probability to be within 15% of the original value and we remove the noisy values. This sometimes leads to missing data points, as explained in How we process the data.

Because attributing searches to regions relies on *general area estimation*, we don't report trends for regions smaller than $3km^2$.

You can learn more about the privacy and quality methods used to generate the data by reading this anonymization process description.

Data fields

The data includes the following fields:

Common 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* or *GB*.
- **sub_region_1**: The name of a region in the country. For example, *Texas* or *Scotland*.
- **sub_region_1_code**: A country-specific ISO 3166-2 code for the region. For example, *US-TX or GB-SCT*.
- **sub_region_2**: The name (or type) of a region in the country. Typically a subdivision of sub_region_1. For example, *Santa Clara County* or *municipal_borough.*
- **sub_region_2_code**: In the US, the FIPS code for a US county (or equivalent). For example, *06085*.
- **sub_region_3**: The name (or type) of a region in the country. Typically a subdivision of sub_region_2. For example, *Downtown* or *postal_code.*
- **sub_region_3_code**: In the US, the ZIP code. In the UK, the postcode district. In Canada, the FSA. In Australia, Postcodes. For example *94303*, *E17**, K1A, or 3000*.**
- **place_id:** The Google place ID for the most-specific subregion. Used in the Google Places API and on Google Maps. For example, *ChIJd_Y0eVIvkIARuQyDN0F1LBA*.

- **date**: The first day of the week (starting on Monday) on which the searches took place. For example, in the weekly data the row labeled *2021-04-19* represents the search activity for the week of April 19 to April 25, 2021, inclusive. Calendar days start and end at midnight Pacific Standard Time, regardless of the region's time zone.

Fields for the 3 main categories

- **sni_covid19_vaccination**: The scaled normalized interest related to all COVID-19 vaccinations topics for the region and date. For example, *87.02*. Empty when data isn't available.
- **sni_vaccination_intent**: The scaled normalized interest related to vaccination intent for the region and date. For example, *22.69*. Empty when data isn't available.
- **sni_safety_side_effects**: The scaled normalized interest related to safety and side effects of the vaccines for the region and date. For example, *17.96*. Empty when data isn't available.

Fields for trending searches (max 20 for each type and category in a region)

- **query_type**: The type of ranking algorithm used to select this group. Either *top* or *rising*.
- **query**: A representative query about Covid-19 vaccinations, which is characteristic of this group of searches. For example, *pfizer vaccine side effects*.
- **rank**: The ranking position for this group, location and time period. 0 is the most frequent search group and 19 the least frequent—when trending searches include a full list of 20 queries. For example, *5*.
- **sni:** The scaled normalized interest in this group for the region and date. For example, *0.34*.
- **history:** A list of the scaled normalized interest values from previous weeks (up to 6). Values are separated by a vertical bar (sometimes called a pipe symbol). For example, *18.0|23.4|50.9.*
- **members:** The 5-most-common member queries of this group. Queries are separated by a vertical bar (sometimes called a pipe symbol). For example, *covid vaccine near me|covid booster near me|get covid booster*.
- **num_members:** How many distinct queries the group contains. For example, *39*.
- **category:** The category this group belongs to. One of *covid19_vaccination*, *vaccination_intent*, or *safety_side_effects*.

Get started

To start working with the vaccination insights data, you can do the following:

1. Explore or download the data using our interactive dashboard.

2. Run queries in Google Cloud's COVID-19 Public Dataset Program.
3. Analyze the data alongside other covariates in the COVID-19 Open-Data repository.

If you download data from the interactive dashboard, the data is organized for each country into the following files:

- **[ISO CODE]_vaccination_search_insights.csv** contains the scaled normalized values using the common and 3-main-category fields detailed above.
- **[ISO CODE]_[REGION LEVEL]_vaccination_trending_searches.csv** files contain the top and rising weekly-trending search topics using the common and trending-search fields detailed above. The region level in the file name corresponds to data for country_region, sub_region_1, and sub_region_2 (not all subregions are available for all countries).

Attribution

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

> Google LLC "Google COVID-19 Vaccination Search Insights". http://goo.gle/covid19vaccinationinsights, Accessed: <date>.

Feedback

We'd love to hear about your project and learn more about your case studies. We'd also appreciate your feedback on the dashboard, data and documentation, or any unexpected results. Please email us at covid-19-search-trends-feedback@google.com.

Availability

We'll continue to update this product while public health experts find it useful in their COVID-19 vaccination efforts. Our published data will remain publicly available to support long-term research and evaluation.

Product changes

Aug 30, 2022 - Included trending-search data in downloads from the dashboard.

Jun 1, 2022 - Added data for Australia.

May 6, 2022 - Added weekly trending searches based on topic clusters (in some countries).

Apr 14, 2022 - Added data for Canada.

Feb 24, 2022 - Added data for Ireland.

Dec 20, 2021 - Added data for the United Kingdom and released trending searches.

Jul 30, 2021 - Documented classifier training and evaluation, anonymization process and categories hierarchy.

Jun 30, 2021 - Public release.