COVID-19 Case Surveillance Public Use Data with Geography Frequently Asked Questions (FAQ)

Date Last Updated: March 22, 2021

This document is intended to provide users of CDC's COVID-19 Case Surveillance Public Use Data with Geography dataset with answers to frequently asked questions. If you have further questions or need support, contact CDC's "Ask SRRG" at eocevent394@cdc.gov.

FAQ Topics

General Information

1) What COVID-19 case surveillance public use datasets are available and how do they differ?

As of March 23, 2021, CDC has three COVID-19 case surveillance datasets for public use:

- **COVID-19 Case Surveillance Public Use Dataset with Geography:** Public use, patient-level dataset with clinical and symptom data, demographics, and state and county of residence. Available on data.cdc.gov. (19 data elements)

- **COVID-19 Case Surveillance Public Use Data:** Public use, patient-level dataset with clinical and symptom data and demographics, with no geographic data. Available on data.cdc.gov. (12 data elements)

- **COVID-19 Case Surveillance Restricted Access Detailed Data:** Restricted access, patient-level dataset with clinical and symptom data, demographics, and state and county of residence. Access requires an easy registration process and a data use agreement. Information available on data.cdc.gov and dataset stored on a secure GitHub repository. (32 data elements)

2) What are the limitations of these data?

For information about national COVID-19 surveillance, how CDC collects and uses COVID-19 surveillance data, the limitations of national case surveillance for COVID-19, or other general COVID-19 data and surveillance questions, visit the FAQ: COVID-19 Data and Surveillance webpage.

3) Why would there be missing information in the COVID-19 Case Surveillance Public Use datasets?

The COVID-19 pandemic has put unprecedented demands on the public health data supply chain. In many states, the large number of COVID-19 cases has severely strained the ability of hospitals, healthcare providers, and laboratories to report cases with complete demographic information, such as race and ethnicity. The unprecedented volume of cases has also limited the ability of state and local health departments to conduct thorough case investigations and collect all case report data. As a result, many COVID-19 case notifications submitted to CDC do not have complete information on patient demographics, clinical outcomes, exposures, and factors that may put people at higher risk for severe disease.

4) Why does CDC provide patient-level datasets to the public?

Sharing timely and accurate COVID-19 data with the public is a core activity of CDC's COVID-19 Emergency Response as well as a key priority of CDC's Data Modernization Initiative and the administration's Executive Order on Ensuring a Data-Driven Response to COVID-19 and Future High-Consequence Public Health Threats. Public datasets are critical for several reasons: open government and transparency, promotion of research, and efficiency (i.e., providing the public, media, and others access to the same data with consistency and supporting information).

5) Are there suggested references on data privacy protections used in the design of these datasets?

- Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data;* March 2007. https://dl.acm.org/doi/10.1145/1217299.1217302.

- Thijs Benschop and Matthew Welch. *Statistical Disclosure Control for Microdata: A Practice Guide for SdcMicro — SDC Practice Guide Documentation*; June 2016. https://sdcpractice.readthedocs.io/en/latest/index.html.

- Centers for Disease Control and Prevention, Agency for Toxic Substances and Disease Registry. (2016). Policy on Public Health Research and Nonresearch Data Management and Access. Centers for Disease Control and Prevention, Department of Health and Human Services; 2016. https://www.cdc.gov/maso/policy/policy385.pdf

- Oleg Chertov and Anastasiya Pilipyuk. Statistical disclosure control methods for microdata. *International Symposium on Computing, Communication and Control*; October 2009. https://www.researchgate.net/publication/228827997_Statistical_Disclosure_Control_Methods_for_Microdata

- Khaled El Emam and Fida Kamal Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*; September- October 2008. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2528029/

- Intergovernmental Data Release Guidelines Working Group (DRGWG) Report. CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data. *CDC Stacks Public Health Publications*; January 2005. https://stacks.cdc.gov/view/cdc/7563

- Richard J. Klein, Suzanne E. Proctor, Manon A. Boudreault and Kathleen M. Turczyn. Healthy People 2010 criteria for data suppression. *Statistical notes*; July 2002. https://www.cdc.gov/nchs/data/statnt/statnt24.pdf

- Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. *Tech. rep SRI-CSL-98-04, SRI Computer Science Laboratory, Palo Alto, CA.* 1998. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.37.5829

- Matthias Templ, Alexander Kowarik and Bernhard Meindl. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software*; October 2015. https://www.jstatsoft.org/article/view/v067i04

6) What privacy protections have been applied to these datasets?

To reduce the risk that these datasets could be used to reidentify persons, CDC designed each dataset accounting for privacy and confidentiality, and conducts ongoing privacy assessments using standard methods and systematically verifies data prior to release. Strict privacy protections, including data suppression, were applied to all three datasets. See the information included with each dataset for more information.

7) Where else can I find COVID-19 data?

COVID-19 data will continue to be made available to the public as summary or aggregate count files, including total counts of cases and deaths by state and by county. These and other data on COVID-19 are available from multiple public locations:

- CDC COVID Data Tracker
- COVID Data Tracker Weekly Review | CDC
- Surveillance & Data Analytics (cdc.gov)
- CDC.gov Data Catalog

8) Where can I learn more about CDC and federal data privacy or open data initiatives?

For more information on CDC open data initiatives, access the CDC COVID-19 Public Health Data Modernization Initiative Fact Sheet or CDC's Data Modernization Initiative. For more on federal open data initiatives, access the 2020 Federal Data Strategy Action Plan.

CDC Support

9) If I have questions, whom should I contact?

Please use the "Contact Dataset Owner" button on the dataset web page or email CDC's Surveillance Review and Response Group (SRRG) at eocevent394@cdc.gov directly.

10) If I want to find more information related to race and ethnicity, what resources does CDC have?

CDC's COVID Data Tracker provides maps, charts, and data on demographic trends among COVID-19 cases by race, ethnicity, age, and sex and urban/rural status and socioeconomic variables. These data are updated daily to provide more timely data, but the public use dataset is designed to not be directly linkable to provide privacy protections.

11) What support do you provide for using these data?

More information is available on the dataset web page, in the data dictionary, and in the developer API docs. You can also ask for help from CDC's Surveillance Review and Response Group (SRRG) by emailing Ask SRRG at eocevent394@cdc.gov.

Technical Information

12) Is the COVID-19 Case Surveillance Public Use Data with Geography dataset in addition to or in replacement of the two current public datasets?

This new dataset is in addition to the existing Public Use (without geographic information) and Restricted Access datasets. Sharing timely and accurate COVID-19 data with the public is a core activity of CDC's COVID-19 Emergency Response and Data Modernization Initiative, and the new administration's Executive Order on Ensuring a Data-Driven Response to COVID-19 and Future High-Consequence Public Health Threats.

13) What is the time period for the COVID-19 Case Surveillance Public Use Data with Geography dataset?

The dataset time period is from January 1, 2020.

14) Are data in the COVID-19 Case Surveillance Public Use Data with Geography dataset provisional?

All COVID-19 case surveillance data are considered provisional by CDC and are subject to change until data are reconciled and verified with jurisdictions. Visit the FAQ: COVID-19 Data and Surveillance webpage for more information.

15) How often will the COVID-19 Case Surveillance Public Use Data with Geography dataset be updated?

The dataset will be updated at the end of each month, following the same process for the current public use dataset with 12 data elements and the restricted access dataset with 32 data elements. The update will replace all prior data to ensure that all data shared by jurisdictions including new cases, corrections, or updates to prior cases, reports of old cases, are included. Record order is randomized each month to reduce the ability to link individual records from each month.

16) What data elements are included in the COVID-19 Case Surveillance Public Use Data with Geography dataset?

The dataset includes 19 data elements, including geographic and demographic variables. Please refer to the dataset web page and the data dictionary for a full list of the variables and their descriptions.

17) Does the COVID-19 Case Surveillance Public Use Data with Geography dataset include all COVID-19 cases?

The dataset will include all cases with the earliest date available in each record (date received by CDC or date related to illness/specimen collection) at least 14 days prior to the creation of the previously updated datasets. This 14-day lag allows case reporting to be stabilized and ensures that time-dependent outcome data are accurately captured. COVID-19 cases with no date are not included.

18) How do counts from the COVID-19 Case Surveillance Public Use Data with Geography dataset relate to other counts published by CDC?

These data are microdata (i.e., individual-person data as opposed to aggregate data) and are shared with CDC by public health jurisdictions and have varying levels of completeness depending on many factors. CDC publishes daily aggregate counts on COVID Data Tracker and data.cdc.gov that are reported by jurisdictions independently from the collection of case data, so counts may differ. CDC publishes the COVID-19 Integrated County View that uses aggregate data using counts only collected from public health jurisdictions that is independent from the collection of these case data, so counts may differ. Data presented may differ from data on state and local websites. This may be due to differences in how data were collected (e.g., date specimen obtained, or date reported for cases) or how the metrics are calculated. Data presented in the county view use standard metrics across all counties in the United States. For the most accurate and up-to-date data for a specific county or state, visit the relevant state or local health department website.

19) What quality assurance procedures are applied?

Data quality assurance procedures (i.e., ongoing corrections and logic checks to address data errors) are performed to ensure quality of the COVID-19 Case Surveillance Public Use Data with Geography dataset. To date, the following data cleaning steps have been implemented:

- Questions that are left unanswered (i.e., blank) on the case report form are re-classified to a missing value, if applicable to the question. For example, in the question, “Was the patient hospitalized?” where the possible answer choices include “Yes,” “No,” or “Unknown,” the blank value is re-coded to missing if the information was not provided by the jurisdiction.
- Logic checks are performed for date-specific data. If an illogical date has been provided, CDC reviews the data with the reporting jurisdiction. For example, if a symptom onset date that is in the future is reported to CDC, this value is set to blank until the reporting jurisdiction updates this information appropriately.

- The case month variable is calculated from the earliest of any clinical date, report date specified by the state health department, or the date the case was received by CDC.

20) We have some counties with very small populations of certain ethnic or racial groups. When talking about other privacy protections, how does that work? (e.g., how are race and ethnicity treated in the dataset?)

One of the reasons we design privacy protection rules are to reduce the risk of identifying individuals based on their demographics and location. For very small populations, this is more important. We remove location information for cases in counties with low populations (i.e., <20,000) and to remove some or all sex, race, and ethnicity information for cases in counties with low subpopulations (i.e., <220) by sex, race, and ethnicity. We enforce a minimum cell size of 11 using all potentially identifying fields; in situations where there are fewer than 11 cases, we suppress one or more fields to ensure the case is contained in a cell with at least 11 cases. The impact of these privacy protections on data should be considered when designing analyses so that incorrect conclusions are avoided concerning prevalence by location and demographics.

21) Is the dataset considered machine readable?

Yes. This dataset was designed following the Findable, Accessible, Interoperable, and Reusable (FAIR) Guiding Principles for scientific data management and requirements of the CDC Data Modernization Initiative and the Federal Data Strategy. The data are available through the Data.CDC.gov web site for people to use with any web browser. The data are also available directly for machines such as automated programs and algorithms created by developers, via the dataset's application programming interface (API), and direct export in open standard formats such as Comma-Separated Values (CSV) and JavaScript Object Notation (JSON). Machine-readable details are available on the dataset's web page. Machine readability is important so that these data can be included and easily be kept up to date and accurate in public health partner and general public web sites, and to help make it easier for the public to include CDC data within the data tools that they wish to use.

22) How can I access the COVID-19 Case Surveillance Public Use Data with Geography dataset?

The data are available to the public on data.cdc.gov in multiple formats, including .csv file, and can be downloaded, accessed via API, or used with online visualization tools. CDC provides a data lens site that allows online exploration with the data. Developer guides are published with examples in many programming languages.

23) Where do the data in the COVID-19 Case Surveillance Public Use Data with Geography dataset come from?

These data are shared with CDC by health departments in states, territories, tribes, and municipalities. CDC is the steward of these data and protects them and organizes them for public health use within the COVID-19 response, including making them available for use by the public.

24) What other datasets might I consider linking to this dataset at the federal level?

By design, geographic linking can be done directly with any other federal datasets that have FIPS codes (e.g., county-level U.S. Census data).

25) I found a field on the case report form that's not in the COVID-19 Case Surveillance Public Use Data with Geography dataset and I want to use it for my research, paper, analysis, or question. How do I get these data?

To protect individual privacy, not all variables on the case report form are released. However, CDC is working to release additional data. You may find that the variable is available in other datasets like the public use dataset or the restricted access dataset. If additional data are released, they will be noted on data.cdc.gov.

Privacy Protection

1) **Is there a lag on the COVID-19 Case Surveillance Public Use Data with Geography dataset while it is being suppressed?**

   No, suppression steps are applied automatically during the generation of the dataset and verified prior to each update.

2) **Will the data be replaced each month or appended?**

   It is a cumulative dataset released every month, so number of cases will be increasing as cases are added. Privacy protection steps are run and verified monthly before release.

3) **Are records removed if field values are suppressed?**

   Records (cases) will not be removed. If field values are suppressed for privacy protection purposes, values will be changed to *NA.* For example, if the total number of cases for a particular state is 1000, then even if county-level information for cases in that state is suppressed to meet privacy rules, the total number of cases in that state would remain 1000.

4) **What is the exact date for each case in the COVID-19 Case Surveillance Public Use Data with Geography dataset? Why do you only release case month? What date is used for case month?**

   Exact dates are not released to protect privacy and reduce the ability to link specific records to other COVID-19 data such as from state or local health department websites. The date used to calculate case month is based on an algorithm that uses the earliest date CDC receives from several data elements including the onset of symptoms, lab specimen, report date specified by jurisdiction, and date CDC receives the case information. This date reflects the closest date to when public health is aware that the case exists and may vary from dates used for other COVID-19 data. To understand time periods related to cases, review the source's information provided for whether the date reflects when the case was referred, when the patient experienced symptom onset, or some other date. Particularly, when looking at cases at the county level, the date used will affect specific counts. The counts will vary depending on location and time period based on factors such as mass testing events, or batches of case results submitted together.

5) **How does suppression affect patterns and trends within the data?**

   Before using data, you should review the level of suppression and patterns within each field to plan your analyses. Because of suppression patterns, data frequency distributions will vary somewhat from actual frequencies. Case report data from smaller populations and rural populations are disproportionately suppressed because counts are more likely to be low and thus more likely to be suppressed by the privacy protection algorithms. A utility summary is created and updated on data.cdc.gov each time the dataset is updated and will show level of completeness and suppression within the data.