Summary, Guidance, and Limitations Information and Restricted Access Data Use Agreement Information

CDC COVID-19 Emergency Response

Centers for Disease Control and Prevention

Suggested Citation: Centers for Disease Control and Prevention, COVID-19 Response. COVID-19 Case Surveillance Restricted Data Access, Summary, and Limitations (dataset access date: Day, Month, Year).

Purpose

The purpose of this document is to facilitate proper access, analysis, and interpretation of two COVID-19 case surveillance datasets that are available to the public—a public use file and a restricted access dataset. The public use file includes 11 data elements with demographic data but no geographical information. It can be downloaded from data.cdc.gov.

The restricted access dataset includes 31 data elements with both state- and county-level information as well as symptom data. The restricted access dataset can be downloaded from a private GitHub account after completing a registration process and data use restrictions agreement.

More information about the registration process, dataset characteristics, case surveillance, and data limitations is provided below.

Introduction

To protect Americans from serious infectious diseases and other health threats, public health authorities conduct national case surveillance to monitor more than 120 diseases and conditions. For these conditions, public health agents collect information on individuals with these infections in a population, which is known as case surveillance. The goal of case surveillance is to provide the information necessary to control outbreaks and inform public health action. Case surveillance is especially important to better understand new diseases, such as COVID-19.

Legislation, regulation, and other rules in jurisdictions require health care providers, hospitals, laboratories, and others to provide information on reportable conditions to public health authorities or their agents by reporting either at the local level (shared up to the state) or directly to the state health department.

COVID-19 is a mandatory reportable condition in all U.S. state health departments, several territorial health departments, and two local health departments (New York City and District of Columbia). These state, territorial, and local health departments determine what information laboratories and health care providers in their areas are asked to collect. The state, territorial and local health departments confirm cases of COVID-19 based on national standardized criteria and may gather additional information on the cases reported. The data elements can be found on the Human Infection with 2019 Novel Coronavirus Case Report Form (CRF). Jurisdictions then voluntarily notify CDC of COVID-19 cases using the National Notifiable Diseases Surveillance System (NNDSS).

Both the public use and restricted access datasets are created from the same data—the COVID-19 case

surveillance notifications shared by jurisdictions with CDC via NNDSS. Differences between the two datasets are described below.

Public Use Dataset

The COVID-19 public use dataset includes 11 data elements and the following variables:

- Initial case report date to CDC
- Date of first positive specimen collection
- Symptom onset date, if symptomatic
- Case status
- Sex
- Age group (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+ years)
- Race and ethnicity (combined)
- Hospitalization status
- ICU admission status
- Mechanical ventilation (MV)/intubation status
- Death status
- Presence of underlying comorbidity or disease

Access the public use file and learn more about this dataset at data.cdc.gov.

Restricted Access Dataset

The restricted access dataset includes 31 data elements and the following variables:

- Initial case report date to CDC
- Date of first positive specimen collection
- Symptom onset date, if symptomatic
- Case status
- Sex
- Age group (0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80+ years)
- Race and ethnicity (combined)
- State of residence
- County of residence
- Healthcare worker status
- Pneumonia present
- Acute respiratory distress syndrome (ARDS) present
- Abnormal chest x-ray (CXR) present
- Hospitalization status
- ICU admission status
- Mechanical ventilation (MV)/intubation status
- Death status
- Presence of each of the following symptoms: fever, subjective fever, chills, myalgia, rhinorrhea, sore throat, cough, shortness of breath, nausea/vomiting, headache, abdominal pain, diarrhea
- Presence of underlying comorbidity or disease

Review the Registration Information and Data Use Restrictions Agreement section in this document for information about accessing this dataset.

Updated on December 14, 2020

Common Dataset Characteristics for Public Use and Restricted Access Datasets

Data are Considered Provisional

- Both the public and restricted access case surveillance datasets are dynamic, meaning case reports can be modified at any time by jurisdictions sharing COVID-19 data with CDC.
- CDC may update prior cases shared with CDC based on any updated information from jurisdictions.
- National case surveillance data are constantly changing. For instance, as new information is gathered about previously reported cases, health departments provide updated data to CDC. As more information and data become available, analyses might find changes in surveillance data and trends during a previously reported time window. Data may also be shared late with CDC due to the volume of COVID-19 cases.
- Annual finalized data: To create the final NNDSS data used in the annual tables, CDC works carefully with the reporting jurisdictions to reconcile the data received during the year until each state or territorial epidemiologist confirms that data from their area are correct.

Version updates to the public and restricted access datasets will be available every 4 weeks. The datasets will include all cases with an initial report date of case to CDC at least 14 days prior to the creation of the previously updated datasets. This month lag will allow adjustments to case reporting and ensure that time-dependent outcome data, including death, are accurately captured.

CDC's Case Surveillance Section routinely performs data quality assurance procedures (i.e., ongoing corrections and logic checks to address data errors). To date, the following data cleaning steps have been implemented:

- Questions that have been left unanswered (blank) on the CRF are re-classified to an *Unknown* 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 missing value is re-coded to the *Unknown* answer option if the respondent did not answer the question.
- Logic checks are performed for date 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 null until the reporting jurisdiction updates this information appropriately.
- The initial report date of the case to CDC is intended to be completed by the reporting jurisdiction when data are submitted. If blank, this variable is completed using the date the data file was first submitted to CDC.
- Additional data quality processing to recode free text data are ongoing. Data on symptoms, race and ethnicity, and healthcare worker status have been prioritized.

Data Suppression for the Public Use and Restricted Access Datasets

To prevent the release of data that could be used to identify persons, data cells are suppressed for low frequency (< 5) records. Records are never removed from the dataset, but individual field values are suppressed for geographic areas with low reporting counts (in the restricted access dataset) or rare combinations of demographic characteristics (sex, age group, race/ethnicity) (in both the restricted access and public use datasets). Suppressed values are re-coded to the *NA* answer option.

Dataset Limitations for the Public Use and Restricted Access Datasets

The COVID-19 case surveillance system is passive, meaning data underestimate the true numbers of cases because of underdiagnosis or underreporting. Completeness of reporting is influenced by many factors (e.g., availability of diagnostic testing, resources, and priorities for health officials). Because reporting to CDC is voluntary, reporting practices vary by state and also depend on a variety of factors. Differences could exist between state-specific databases and CDC's COVID-19 surveillance database, though efforts are made to align CDC's database with state-specific data.

Although the CRF captures several outcomes, including hospitalization, ICU admission, and death, these data may be incomplete because outcomes are not yet known at the time of reporting (i.e., outcomes coded as *Unknown*). These data elements also may not represent final outcomes, as a patient's condition may have changed after case data submission, but the case report was not updated.

Registration Information and Data Use Restrictions Agreement

COVID-19 Case Surveillance Restricted Access Detailed Data
Registration Information and Data Use Restrictions Agreement (RIDURA)

After reviewing the information below, use the electronic form to request access to the restricted access dataset.

Data Use Restrictions Agreement

Upon submission of this request for access and granting of that access to the data, I attest and agree to comply with the following terms:

Security

1. I understand and agree to the following security practices:
    a. All listed requesters must use appropriate safeguards to protect the data from misuse or inappropriate disclosure and prevent any use or disclosure of the data other than as provided in this RIDURA or as otherwise required by law.
    b. I will password protect the restricted access data provided herein.
    c. I will treat the restricted access data provided herein confidentially and will not give other persons access, other than co-requesters, unless otherwise required by law.
    d. Any hard copies of data will be kept in a locked office cabinet, with access limited only to the primary requester and any co-requesters.
    e. The primary requester must report any loss or misuse of data to the CDC (eocevent394@cdc.gov) within three (3) business days after the loss or misuse is discovered.
    f. Data will not be transmitted between computer systems, or via email or email attachment, unless the transmission uses Secure Socket Layer (SSL) RC4 128 bit algorithms, SSL Server-Gated Cryptography (SGC) 128 bit algorithms, TLS 1.11 128 bit algorithms, or other algorithms accepted and certified by the National Institute of Standards and Technology.
    g. The Requester agrees to maintain, store, protect, archive and/or dispose of data in accordance with applicable law.

Access and Use

2. I am responsible for obtaining Institutional Review Board review of projects when appropriate.
3. Access and use of the data and/or information does not grant me permission to use any trade names, trademarks, services marks, product names, or logos of CDC or the Department of Health and Human

Services, except as may be required for reasonable and customary use in describing the CDC or the data and/or information. I will obtain express written approval from CDC prior to any use of the. Though I agree to identify CDC as the source of the data provided, I further agree to not imply or state in any written form, that use of or any interpretation based on the data are those of the original data sources or of CDC.

4. I understand that use of these data does not imply endorsement by CDC. I will not attribute any analysis conducted using these data to CDC.
5. I agree that while matching cases for public health purposes is acceptable, I will not deliberately participate in or support the combination of case surveillance datasets with other datasets for the specific purpose of matching records to identify individuals.
6. I understand that CDC has taken all reasonable steps for privacy protections to ensure the identity of data subjects cannot be disclosed. No direct identifiers or characteristics that might lead to identification have been included in the data provided. As such, I will not use the data to re-identify or attempt to re-identify any individual included in the data and will not use, publish or release the data in any personally identifiable form. Should I inadvertently re-identify an individual, I will notify CDC of such re-identification within three (3) days of any such discovery.

Presentations, Publications and Dissemination

7. CDC requests a copy of any presentations, publications, or other material shared with the public, sent no later than 4 weeks post-publication/event to eocevent394@cdc.gov.
8. CDC does not warrant that the data and/or information will meet my requirements and disclaims all other warranties and conditions either expressed or implied, including the warranty of merchantability and fitness for a particular purpose.
9. All publications and/or presentations using the restricted access data must include the following disclaimer: "The CDC does not take responsibility for the scientific validity or accuracy of methodology, results, statistical analyses, or conclusions presented."
10. For oral or written presentations or publications, the source of the data must be attributed to the CDC: “Centers for Disease Control and Prevention, COVID-19 Response. COVID-19 Case Surveillance Data Access, Summary, and Limitations (*version date*)" (Please check the GitHub project to include the version date of the dataset that you used for the publication.)

This Agreement is governed by applicable federal law.

Submit the following information using the electronic form, https://forms.gle/pCvbosyaoqRvQmeb7, to start your access request for the COVID-19 Case Surveillance Restricted Access Dataset. Please note that CDC may use the following: affiliation and proposed use of the data information in metrics reporting on users of this dataset.

Information Required:

Primary Requester

Email address
Name
Title
Affiliation
Affiliation Mailing Address
Telephone

What GitHub Account ID do you plan to use?

Updated on December 14, 2020

Your GitHub ID will be granted access to a private repository containing data that we use to make it easier to share data with you. If you do not have an ID, you can create one for free at GitHub.com. After review, you will receive an email invitation from a CDC staff member. Please email eocevent394@cdc.gov if you have any questions.

Proposed use of the data

- Title of Analysis
- Brief description of proposed analysis
- Purpose of analysis / Public health significance
- Describe the intended products from this analysis

Note: If other individuals are working on any analyses with the primary requester, please provide their information as co-requesters. If the primary requester is a trainee, student, intern, fellow or requesting the data for use in any type of training program, please provide the primary requester's supervisor as a co-requester.

Co-Requesters

- Email address
- Name
- Title
- Affiliation
- Affiliation Mailing Address

Additional COVID-19 Data

COVID-19 data will 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, including:

https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html

https://www.cdc.gov/covid-data-tracker/index.html

https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html

https://www.cdc.gov/coronavirus/2019-ncov/php/open-america/surveillance-data-analytics.html

Questions

The Case Surveillance Task Force and Surveillance Review and Response Group (SRRG), part of CDC's COVID-19 Emergency Response, are the stewards for the restricted access dataset. If you have questions about the restricted access dataset, contact ASK SRRG (eocvent394@cdc.gov). Information about the restricted access dataset is also available on CDC's website at https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t.

Updated on December 14, 2020