COVID-19 Case Surveillance Data Access, Summary, Guidance, and Limitations

Case Surveillance Task Force, CDC COVID-19 Response, May 2020

U.S. Centers for Disease Control and Prevention

Suggested Citation: Centers for Disease Control and Prevention, COVID-19 Response. COVID-19 Case Surveillance Data Access, Summary, and Limitations (version date: May 18, 2020).

Purpose

The purpose of this document is to facilitate proper access, analysis, and interpretation of the novel coronavirus (COVID-19) case surveillance data. The document summarizes important information on the data access process and describes limitations of the case surveillance data.

Introduction

The COVID-19 case surveillance system database includes patient-level data reported to U.S. states and autonomous reporting entities, including New York City and the District of Columbia (D.C.), as well as U.S. territories and states. On April 5, 2020, COVID-19 was added to the *Nationally Notifiable Condition List* and classified as "immediately notifiable, urgent (within 24 hours)" by a Council of State and Territorial Epidemiologists (CSTE) Interim Position Statement (Interim-20-ID-01). The statement also recommended that all states and territories enact laws to make COVID-19 reportable in their jurisdiction, and that jurisdictions conducting surveillance should submit case notifications to CDC. COVID-19 case surveillance data are collected and reported voluntarily to CDC's COVID-19 Response. These data include demographic characteristics, exposure history, disease severity indicators and outcomes, clinical data, laboratory diagnostic test results, and comorbidities. All data elements can be found on the COVID-19 case report form located at www.cdc.gov/coronavirus/2019-ncov/downloads/pui-form.pdf.

Data Access Process

The Case Surveillance Task Force and Surveillance Review and Response Group (SRRG) within CDC's COVID-19 Response provide stewardship for datasets that support the public health community's access to COVID-19 data while protecting patient privacy. Data are made available for limited use upon completion of the registration information and data use restrictions agreement (RIDURA).

- To initiate a request, please use the following "ASK SRRG" email address to contact data stewards: eocevent394@cdc.gov.
- SRRG will provide further documentation and guidance under the RIDURA to access and use COVID-19 case surveillance data appropriately.
- Data requests will be prioritized based a clear description of the immediate impact for COVID-19 response that is anticipated to follow from data use.
- Access will be granted to the appropriate single repository containing data files and data dictionary through the https://github.com/cdc-data

Data Release Specifications

Two line-listed datasets of all COVID-19 cases reported to CDC are available. Either dataset, but not both, is to be made available for limited use upon completion of the RIDURA. COVID-19 data may differ in the variables reported and in completeness by state.

Data by county or any geographic unit smaller than state cannot be released. If the total number of cases in a state in one year is <5 cases, state will be suppressed for all cases in that state.

Dataset 1: *Detailed* (filename: COVID_Cases_Restricted_Detailed_*date*.csv) does not include state of residence.
Variables:

- Initial report date of case to CDC
- Date of first positive specimen collection
- Case status
- Sex
- Age group (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+ years)
- Race
- Ethnicity
- Healthcare worker status
- Symptom onset date, if symptomatic
- 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

Dataset 2: *Limited* (filename: COVID_Cases_Restricted_Limited_*date*.csv) includes state of residence.
Variables:

- Initial date report of case to CDC
- Specimen collection date
- Case status
- Sex
- Age group (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+ years)
- State of residence
- Hospitalization status
- ICU admission status
- Death status
- Underlying comorbidity or disease present

Case Data Standardization

COVID-19 case reports have been routinely submitted using standardized case reporting forms. On April 5, 2020, CSTE released an Interim Position Statement with national surveillance case definitions for COVID-19 included. Current versions of these case definitions are available here: https://wwwn.cdc.gov/nndss/conditions/coronavirus-disease-2019-covid-19/. All cases reported on or after were requested to be reported by public health departments to CDC using the standardized case definitions for lab-confirmed or probable cases. On May 5, 2020, the standardized case reporting form was revised. Implementation of case reporting using this new form is ongoing among U.S. states and territories.

Dataset Versions and Release Schedule

The COVID-19 case surveillance data are dynamic; case reports can be modified at any time by the reporting jurisdiction as new information becomes available (i.e., data are subject to change). Furthermore, reporting jurisdictions may report cases late. Version updates to the detailed and limited datasets will be available for request once a month. The datasets will include all cases with an initial report date of case to CDC at least 30 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. Releases will be managed through github.com and will contain most recent and previous versions (https://github.com/cdc-data).

CDC's Case Surveillance Task Force 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 case report form 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

To prevent release of data that could be used to identify persons, data cells are suppressed for low frequency (<5) records. Suppression includes states with low reporting counts and uncommon combinations of demographic characteristics (sex, age group, race/ethnicity). Suppressed values are re-coded to the *Unknown* answer option.

Dataset Limitations

The COVID-19 case surveillance system is passive; 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 health officials). Because reporting 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 case report form 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.

Data Requests from Agencies, Institutions, or Persons Outside the COVID-19 CDC EOC Response, Including Other CDC Employees

There will be no release of data in formats other than those described above, unless the format is more restrictive than described above. Requests for data must be made using the form, *Registration Information and Data Use Restrictions Agreement (RIDURA)*. Any agency, institution, or person (including other federal agencies) seeking more detailed data than available in the data sets described above will be directed by the response to each state, so that data requestors can negotiate data release and obtain data directly from individual states.

Alternative Methods of Access to Summary 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:

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