Menu

AT

Analysis & updates > Federal COVID Data 101

Federal COVID Data 101: Working with Case Data

Here’s a walkthrough of the Centers for Disease Control and Prevention (CDC) daily case dataset and what you should know about it.

By Jennifer Clyde
February 24, 2021

As of **March 7, 2021** we are no longer collecting new data. Learn about available federal data.

Before COVID-19, public health departments were already reporting nationally notifiable disease case data (i.e., diseases that are required by law to be reported to government authorities for the purpose of disease monitoring and outbreak detection) to the CDC using the National Notifiable Diseases Surveillance System. However, the novel SARS-CoV-2 virus and the urgency of a global pandemic created the need for *daily* counts of cases and deaths that could be reported publicly on a national level. To produce a more expedient dataset, the CDC created a “robust, multistep process to collect data and confirm the case and death numbers with jurisdictions daily.”

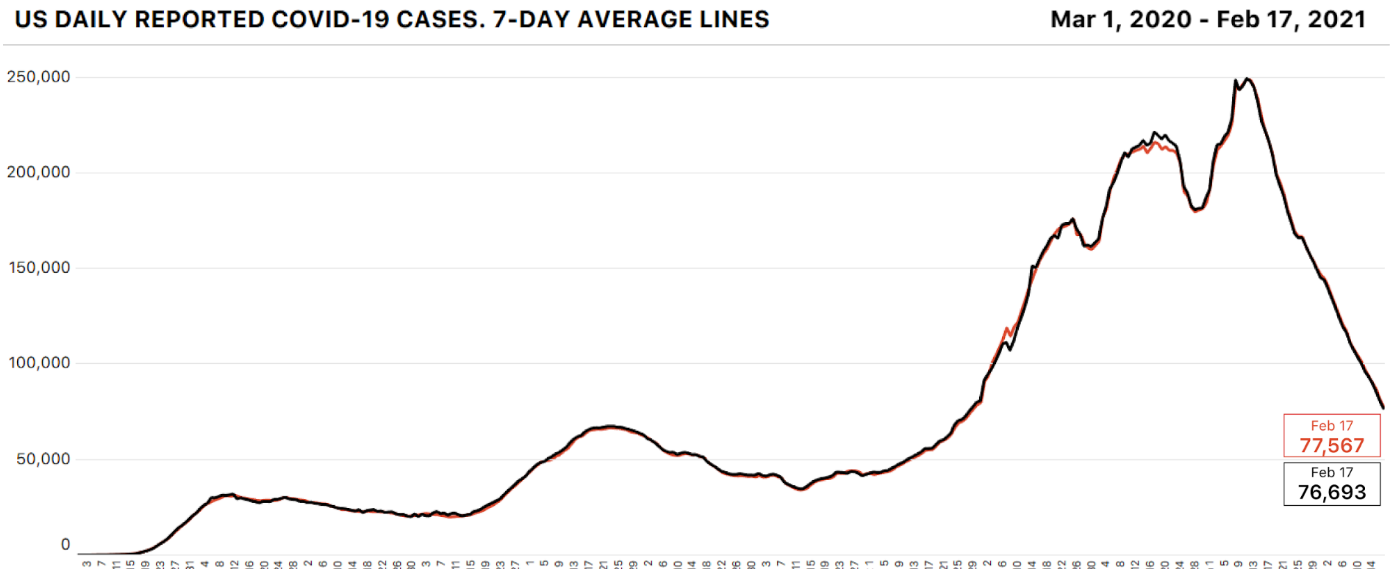
The dataset has included both confirmed and probable cases of COVID-19 since

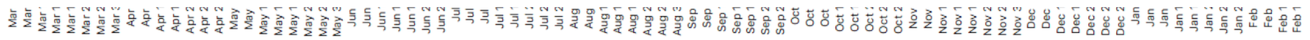
The Council for State and Territorial Epidemiologists issued an April 5, 2020, Interim Case Definition (revised August 5th 2020) in which they noted that “field investigations will involve evaluations of persons with no symptoms and these individuals will need to be counted as cases” and then laid out the criteria for determining a confirmed or probable case. We won’t discuss case definitions in detail here, but will write more about them in an upcoming post.

Two levels of detail

The CDC publishes two main types of case data: aggregate and line level. **Aggregate data** includes the total numbers of cases by jurisdiction, as well as totals for confirmed and probable cases where available. **Line-level data** includes a de-identified line for each case with detailed demographic data. The aggregate data is updated daily, while the line-level data is only updated monthly due to its complexity. We’re going to cover the aggregate dataset (United States COVID-19 Cases and Deaths by State over Time) and only touch briefly on the line-level data, which is generally used in advanced epidemiological research.The sections of the dataset that cover deaths data will be addressed in a separate post in our Federal COVID Data 101 series.

How close are the numbers? Very close.





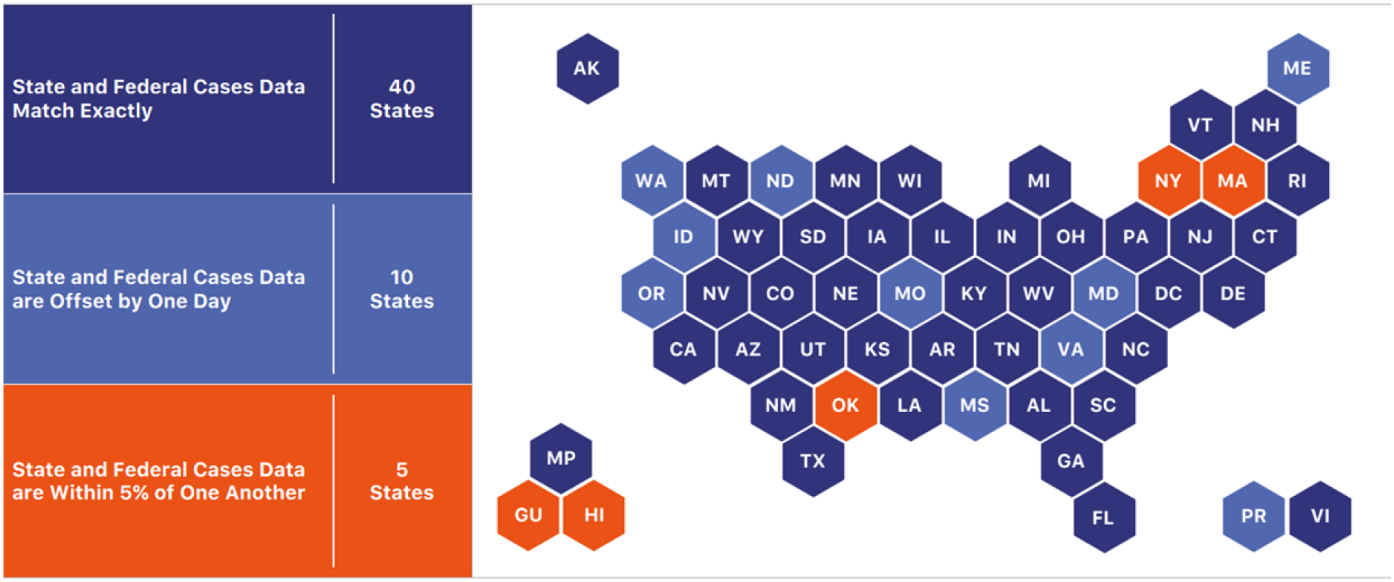
Source: The COVID Tracking Project & HealthData.gov

🟥 COVID Tracking Project ⬛ CDC

Our comparisons of COVID Tracking Project and CDC data for the 56 US states and territories we capture data for show that the **CDC cases data is extremely similar to the state-reported data we have been capturing.** (The CDC data also captures the additional jurisdictions of New York City, and three independent countries in compacts of free association with the United States.)[1]

In our comparison analysis, we found that the total cases in 40 states match exactly; 10 additional states match exactly when offset by one day (i.e. the CDC data is one day ahead of the state reported data), and the remaining six jurisdictions only deviate by up to 5 percent.[2] In addition to these matches at the total cases level, 25 states and territories exactly match across all three metrics we compared: total, confirmed, and probable cases.

COVID TRACKING PROJECT CASES DATA VS FEDERAL CASES DATA


State and Federal Cases Data Match Exactly
40 States
State and Federal Cases Data are Offset by One Day
10 States
State and Federal Cases Data are Within 5% of One Another
5 States
AK
ME
VT
NH
WA
MT
ND
MN
WI
MI
NY
MA
RI
ID
WY
SD
IA
IL
IN
OH
PA
NJ
CT
OR
NV
CO
NE
MO
KY
WV
MD
DC
DE
CA
AZ
UT
KS
AR
TN
VA
NC
NM
OK
LA
MS
AL
SC
MP
TX
GA
GU
HI
FL
PR
VI

Source: The COVID Tracking Project & HealthData.gov

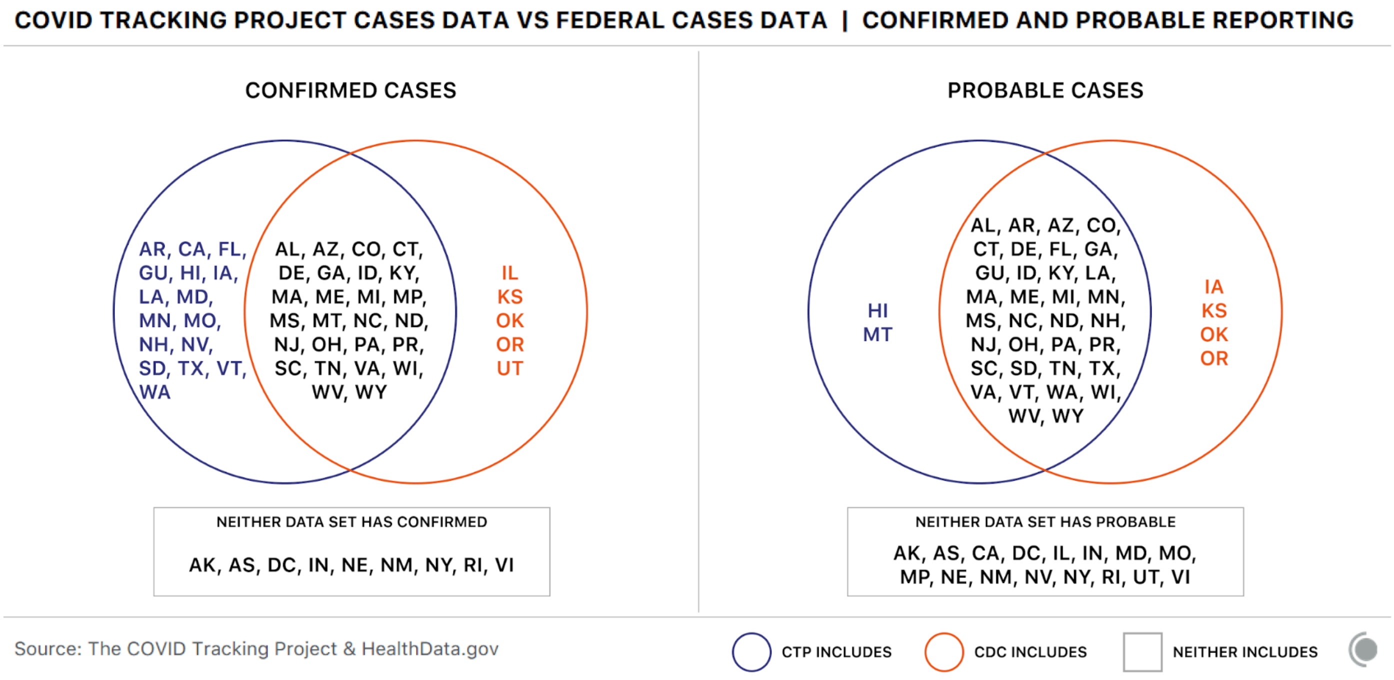
American Samoa not pictured (fewer than 5 total cases)

Of the states and territories that do not match exactly with or without a one-day offset, three (Guam, New York, and Oklahoma) differ by one percent or less—differences likely caused by slight timing variations in data collection, or differences in backfills done by the state versus those done by the CDC, or both. We have identified the source of the differences in the remaining three jurisdictions: Hawaii's probable cases are not included in the CDC data; probable case numbers for Massachusetts have been frozen in the CDC data since September 7, 2020; and the CDC is counting four cases in American Samoa, while we have not counted these cases (from what we can tell they were all sailors who came to American Samoa, and were not allowed to leave their ships). We have

notified these jurisdictions of our findings, and we hope that they will be able to resolve these differences between the state and federal data.

Which case metrics are reported in the two datasets

States and territories can report up to three case metrics: total cases, confirmed cases, and probable cases. Every jurisdiction reports total cases in both state and CDC data, but confirmed and probable cases are not always reported in the same way on state dashboards as they are by the CDC. For 34 states or territories, the same case metrics—although not necessarily the same case *numbers*—appear in the state and federal data. In the remaining states and territories, confirmed cases are reported differently in state and federal reporting for 21 states, while probable cases are reported differently in state and federal reporting for only six states.


COVID TRACKING PROJECT CASES DATA VS FEDERAL CASES DATA | CONFIRMED AND PROBABLE REPORTING
CONFIRMED CASES
AR, CA, FL, GU, HI, IA, LA, MD, MN, MO, NH, NV, SD, TX, VT, WA
AL, AZ, CO, CT, DE, GA, ID, KY, MA, ME, MI, MP, MS, MT, NC, ND, NJ, OH, PA, PR, SC, TN, VA, WI, WV, WY
IL KS OK OR UT
NEITHER DATA SET HAS CONFIRMED
AK, AS, DC, IN, NE, NM, NY, RI, VI
PROBABLE CASES
HI MT
AL, AR, AZ, CO, CT, DE, FL, GA, GU, ID, KY, LA, MA, ME, MI, MN, MS, NC, ND, NH, NJ, OH, PA, PR, SC, SD, TN, TX, VA, VT, WA, WI, WV, WY
IA KS OK OR
NEITHER DATA SET HAS PROBABLE
AK, AS, CA, DC, IL, IN, MD, MO, MP, NE, NM, NV, NY, RI, UT, VI
CTP INCLUDES
CDC INCLUDES
NEITHER INCLUDES

Source: The COVID Tracking Project & HealthData.gov

It’s possible many of these differences will be definitional: that is, states and territories may refer to cases as “confirmed” or “probable” in their public reporting, but those definitions don’t always match the CSTE criteria mentioned above. It’s also possible that it has nothing to do with the definitions, and is a problem in the reporting process—since we don’t have direct access to the reporting from state governments to the federal government, we can’t ascertain what might be happening there. The good news is the CDC does have access to that back-end data stream, and there are encouraging signs that the federal government is working toward releasing more data and more metadata. In some cases, the CDC data already provides more information than the state-reported data, such as in Kansas, where the state only posts a total case number, but the CDC publishes a confirmed and probable case breakdown.

Additionally—and importantly for data users—in 11 states or territories, the CDC isn’t reporting probable cases in the expected data field. The CDC dataset contains two columns related to probable cases: `prob_cases` defined as the “total probable cases” and the `pnew_case` field defined as the “number of new probable cases.” We would expect to see values in `pnew_case` only when values are also present in `prob_cases` . However in Arkansas, Florida, Guam, Iowa, Louisiana, Minnesota, New Hampshire, South Dakota, Texas, Vermont, and Washington, probable values are only available in `pnew_case` . In each of these states the sum of `pnew_case` matches the total of probable cases reported by the state.[3]

Based on our conversations with officials in some of these states, it appears that for the majority—or perhaps all—of the states in this category, the data is reported in this way because those states did not initially grant the CDC permission to publicly report their confirmed and probable case breakdown, and then did not update their permissions when they started reporting confirmed and probable breakdowns in public. **Again, the good news here is that this can be fixed quite easily with a little more coordination between state and federal governments.**

What’s missing

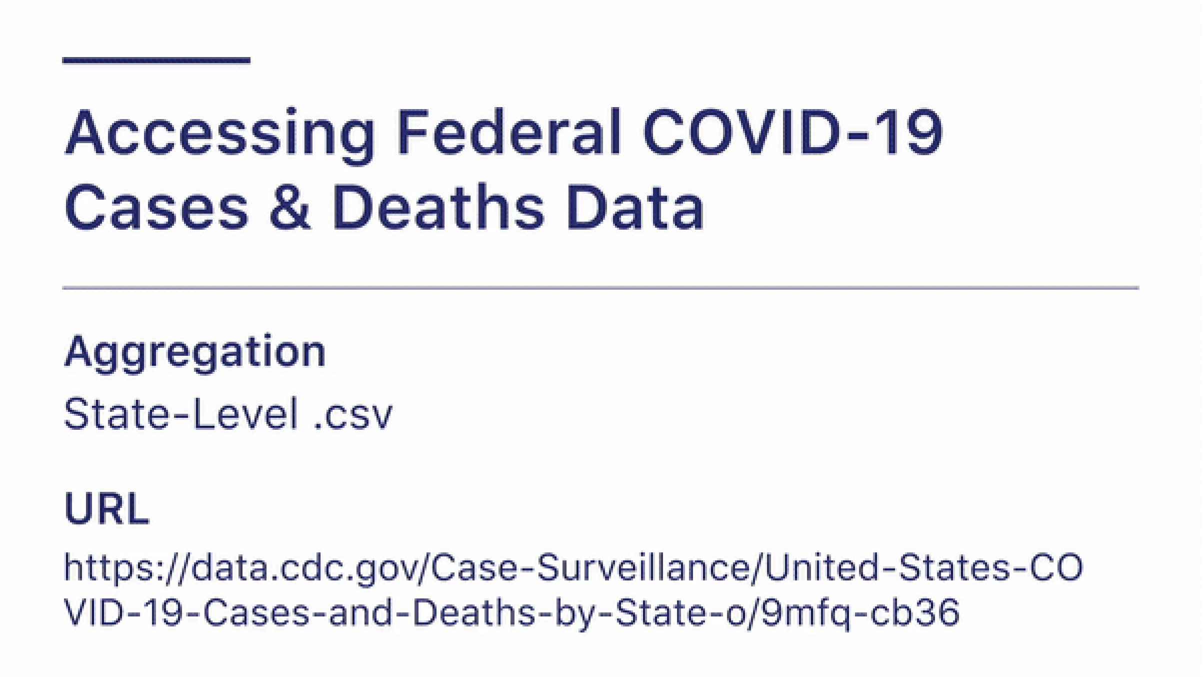
One notable absence in this dataset is the lack of data notes and revision history. Data notes that address anomalies (unexpected decreases, data dumps, etc.) in the case data can be found in the downloadable xlsx files of the Community Profile Reports, however the notes are not available in a structured format. (For our purposes, “structured” means that each note is associated with the appropriate state and date(s) in fixed fields so that data users can use this information either automatically or manually.)

Adding structured notes to the dataset would make them easily accessible for data users seeking to contextualize data. It is also currently unclear whether backfills or corrections are ever included in the existing (unstructured) notes, though of course they do happen in the data; we would recommend including these in the notes. We also recommend that the federal government offer a clear record of revisions using a publicly accessible method, which would enhance transparency and help build trust in the federal data.

Where to find, and how to use, this data

This aggregate cases and deaths dataset is used in the COVID Data Tracker, COVID Data Tracker Weekly Review, Community Profile Reports and State Profile Reports. It is available for download both from the CDC and on HealthData.gov, in a variety of formats including CSV and XML. The CSVs are straightforward and easy to work with.

It is also possible to filter, sort, and visualize the data on the CDC website without downloading it. You can also query the data online via the Socrata Open Data API after consulting the excellent and comprehensive documentation provided.



Accessing Federal COVID-19 Cases & Deaths Data

Aggregation

State-Level .csv

URL

https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

In the dataset, each row corresponds to one state for one day going back to January 2020; the data is updated twice daily, and major disruptions in the data are included in the Community Profile Reports. The data dictionary provided by the CDC is fairly minimal, but clear and adequate for a dataset with only 15 data fields.

CDC Description

Date of counts

CDC Column Name

```
submission_date
```

CTP API Field Name

`date`

CDC Description

Date and time record was created

CDC Column Name

`created_at`

CTP API Field Name

`lastUpdatedEt`

CDC Description

Total number of cases

CDC Column Name

`tot_cases`

CTP API Field Name

`positive`

CDC Description

Total confirmed cases

CDC Column Name

`conf_cases`

CTP API Field Name

`positiveCasesViral`

CDC Description

Total probable cases

CDC Column Name

`prob_cases`

CTP API Field Name

`probableCases`

CDC Description

Number of new cases

CDC Column Name

CDC Column Name

`new_case`

CTP API Field Name

`positiveIncrease` *

CDC Description

Number of new probable cases

CDC Column Name

`pnew_case`

CTP API Field Name

N/A

CDC Description

Total number of deaths

CDC Column Name

`tot_death`

CTP API Field Name

`death`

CDC Description

Total number of confirmed deaths

CDC Column Name

```
conf_death
```

CTP API Field Name

```
deathConfirmed
```

CDC Description

Total number of probable deaths

CDC Column Name

```
prob_death
```

CTP API Field Name

```
deathProbable
```

CDC Description

Number of new deaths

CDC Column Name

`new_death`

CTP API Field Name

`deathIncrease` *

CDC Description

Number of new probable deaths

CDC Column Name

`pnew_death`

CTP API Field Name

N/A

CDC Description

Jurisdiction

CDC Column Name

`state`

CTP API Field Name

`state`

CDC Description

If Agree, then confirmed and probable cases are included. If Not Agree, only total cases.

CDC Column Name

`consent_cases`

CTP API Field Name

N/A

CDC Description

If Agree, then confirmed and probable deaths are included. If Not Agree, only total deeaths.

CDC Column Name

`pnew_death`

CTP API Field Name

N/A

* `positiveIncrease` and `death increase` represent the difference from previous day topline metrics rather than actual new cases/deaths on a given day

For advanced users, there is the previously mentioned line-level data available in two different datasets: the COVID-19 Case Surveillance Public Use Data and the COVID-19 Case Surveillance Restricted Access Detailed Data, which—as the name suggests—has to be requested from the CDC and used for specific purposes only. The primary difference between the two datasets is the level of detail: the public use dataset contains 12 additional data fields and the restricted data set includes 32 additional fields. Additional information including FAQs and the data dictionaries can be found linked from the dataset pages.

The upshot

Overall, the aggregate federal data is a very good replacement for state reported data, and the divergences we've described are not significant blockers to using the aggregate federal dataset now. We'd like to see the federal government provide relevant data notes in a structured way with the dataset, which will allow users to work with the data more responsibly and transparently. Additionally, we hope that going forward, the federal and state governments are able to work more closely together to reconcile the differences in confirmed and probable reporting. But even with these caveats, **the dataset is an excellent successor for COVID Tracking Project's state-reported case data**.

Additional research and contributions from Michal Mart, Dave Luo, and Peter Walker

[1] For comparison purposes, we added the New York state and city totals together in the CDC data.

[2] In Iowa, where the CTP data does not capture probable cases, there appears to be a large gap between the CDC total cases and CTP total cases. Iowa's probable cases are currently captured in our data in the "positive antigen tests people" field because we're missing multiple public time series for Iowa's COVID-19 metrics that would be required to move these figures into the "probable cases" field. When we interpret the figure for Iowa's positive antigen tests as probable cases, the case data provided by Iowa on its dashboard matches perfectly to the Iowa data in the CDC dataset.

[3] We are counting the 11 states described here as reporting probable cases only in the CDC's `pnew_case` field as reporting probable cases to the CDC.

Jennifer Clyde is a data quality and data entry shift lead at the COVID Tracking Project.

More "Federal COVID Data 101" posts

Federal COVID-19 Testing Data Is Getting Better

The federal government improved its state and county-level COVID-19 PCR testing data since we analyzed it in February. Here's an update on those changes and what we hope to see next for the data.

By Kara W. Schechtman
April 21, 2021

Federal COVID Data 101: What We Know About Race and Ethnicity Data

Publicly available federal race and ethnicity COVID-19 data is currently usable and improving, although it shares many of the problems we've found in state-reported data.

By Alice Goldfarb
March 19, 2021

How Not to Interpret COVID-19 Data

Beware of dating schemes, data dumps, weather events and other issues that can lead to mistakes that confuse the public.

By Erin Kissane & Jessica Malaty Rivera
April 1, 2021

See all analysis & updates →

Contact

RSS

About Us

Terms and Conditions

License

Privacy Policy

Accessibility

Sitemap

The COVID Tracking Project collects and publishes the most complete data about COVID-19 in the US.

CovidTracking.com Copyright © 2021 by The Atlantic Monthly Group. (CC BY 4.0)