GAO

United States Government Accountability Office

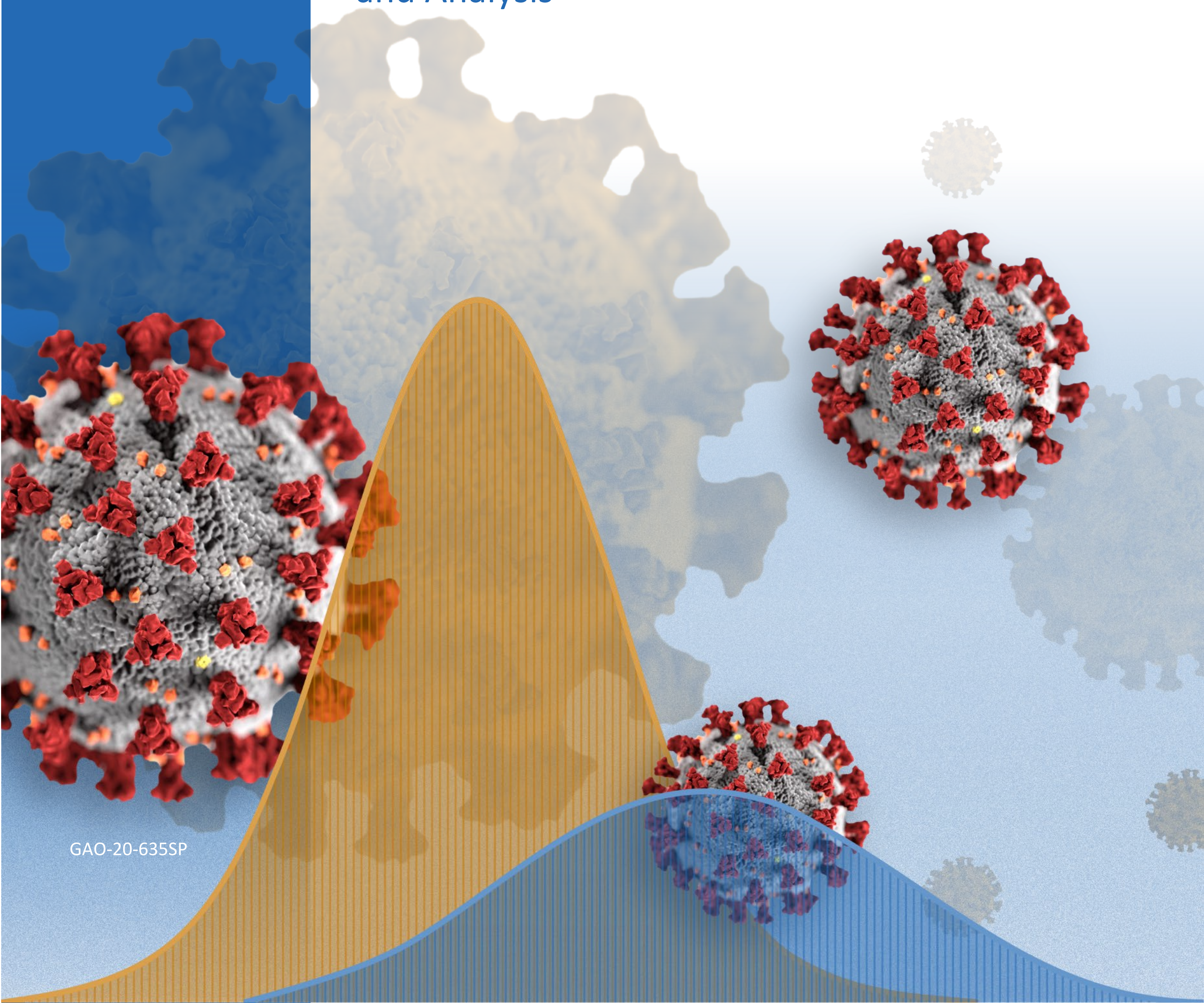
Report to Congressional Committees

July 2020

TECHNOLOGY ASSESSMENT

COVID-19

Data Quality and Considerations for Modeling and Analysis



GAO-20-635SP

The cover displays images of the SARS-CoV-2 virus that causes COVID-19, overlaid with a graph illustrating the concept of “flattening the curve.” Flattening the curve is a term used to describe reducing or delaying the spread of a disease through interventions such as social distancing. Source: Centers for Disease Control and Prevention, Alissa Eckert, Dan Higgins, and GAO analysis.

Highlights of GAO-20-635SP, a report to congressional committees

July 2020

Why GAO did this study

The COVID-19 pandemic has resulted in significant loss of life and profoundly disrupted the U.S. economy and society, and the Congress has taken action to support a multifaceted federal response on an unprecedented scale. It is important for decision makers to understand the limitations of COVID-19 data, and the uses and limitations of various methods of analyzing and interpreting those data.

The Coronavirus Aid, Relief, and Economic Security Act (CARES Act) includes a provision for GAO to, in general, conduct monitoring and oversight of the authorities and funding provided to address the COVID-19 pandemic and the effect of the pandemic on the health, economy, and public and private institutions of the U.S. This technology assessment examines (1) collection methods and limitations of COVID-19 surveillance data reported by CDC, (2) approaches for analyzing COVID-19 data, and (3) uses and limitations of forecast modeling for understanding COVID-19.

In conducting this assessment, GAO obtained publicly available information from CDC and state health departments, among other sources, and reviewed relevant peer reviewed and preprint (non-peer-reviewed) literature, as well as published technical data on specific models.

View GAO-20-635SP. For more information, contact Timothy M. Persons, PhD, at (202) 512-6888 or personst@gao.gov, SaraAnn Moessbauer at (202) 512-4943, or MoessbauerS@gao.gov, or Mary Denigan-Macauley, PhD, at (202) 512-7114 or DeniganMacauleyM@gao.gov.

TECHNOLOGY ASSESSMENT

COVID-19

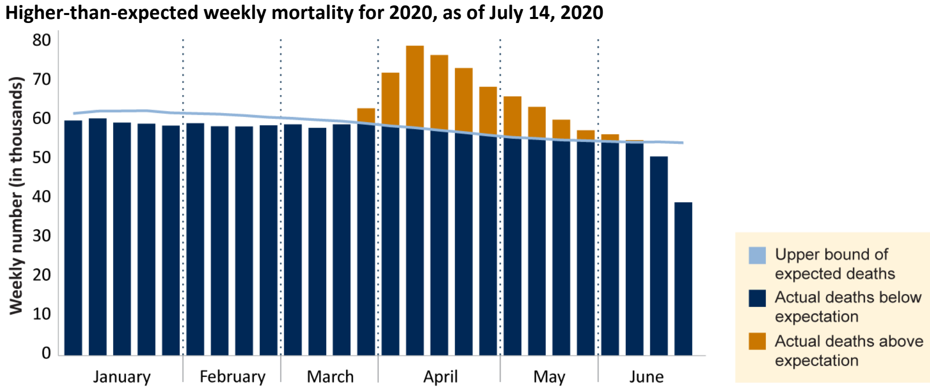
Data Quality and Considerations for Modeling and Analysis

What GAO found

The rapid spread and magnitude of the COVID-19 pandemic have underscored the importance of having quality data, analyses, and models describing the potential trajectory of COVID-19 to help understand the effects of the disease in the U.S. The Centers for Disease Control and Prevention (CDC) is using multiple surveillance systems to collect data on COVID-19 in the U.S. in collaboration with state, local, and academic and other partners. The data from these surveillance systems can be useful for understanding the disease, but decision makers and analysts must understand their limitations in order to interpret them properly. For example, surveillance data on the number of reported COVID-19 cases are incomplete for a number of reasons, and they are an undercount the true number of cases, according to CDC and others.

There are multiple approaches to analyzing COVID-19 data that yield different insights. For example, some approaches can help compare the effects of the disease across population groups. Additional analytical approaches can help to address incomplete and inconsistent reporting of COVID-19 deaths as well. For example, analysts can examine the number of deaths beyond what would normally be expected in the absence of the pandemic. Examining higher-than-expected deaths from all causes helps to address limitations in the reporting of COVID-19 deaths because the number of total deaths is likely more accurate than the numbers of deaths from specific causes. The figure below shows actual deaths from the weeks ending January 1 through June 27, 2020, based on data from CDC's National Center for Health Statistics, compared with the expected deaths based on prior years' data. Deaths that exceeded this threshold starting in late March are considered excess deaths that may be related to the COVID-19 pandemic.

Higher-than-expected weekly mortality for 2020, as of July 14, 2020



Source: GAO analysis of Centers for Disease Control and Prevention (CDC)/National Center for Health Statistics (NCHS) data. | GAO-20-635SP

Analysts have used several forecasting models to predict the spread of COVID-19, and understanding these models requires understanding their purpose and limitations. For example, some models attempt to predict the effects of various interventions, whereas other models attempt to forecast the number of cases based on current data. At the beginning of an outbreak, such predictions are less likely to be accurate, but accuracy can improve as the disease becomes better understood.

Table of Contents

Abbreviations

CARES Act	Coronavirus Aid, Relief, and Economic Security Act
CDC	Centers for Disease Control and Prevention
COVID-19	Coronavirus Disease 2019
COVID-NET	COVID-19-Associated Hospitalization Surveillance Network
HHS	Department of Health and Human Services
IHME	Institute for Health Metrics and Evaluation
NCHS	National Center for Health Statistics
NHSN	National Healthcare Safety Network
SARS	Severe Acute Respiratory Syndrome
SARS-CoV-2	Severe Acute Respiratory Syndrome, coronavirus 2
SEIR	Susceptible-Exposed-Infectious-Recovered

U.S. GOVERNMENT ACCOUNTABILITY OFFICE
441 G St. N.W.
Washington, DC 20548

July 30, 2020

Congressional Committees

The Coronavirus Disease 2019 (COVID-19) pandemic has resulted in significant loss of life and profoundly disrupted the U.S. economy and society, and it has elicited a federal response on an unprecedented scale. This disease is caused by SARS-CoV-2 (Severe Acute Respiratory Syndrome, coronavirus 2), a strain of coronavirus to which the public does not have immunity. It was first reported on December 31, 2019, in Wuhan, China, and in the weeks that followed, it quickly spread around the world, including to the United States. In the absence of medical countermeasures, the United States is using social distancing to slow the spread of the virus.[1] The Centers for Disease Control and Prevention (CDC) is responsible for gathering and reporting data on cases, hospitalizations, and deaths due to COVID-19. Quality data are paramount for understanding how the virus is affecting the population. It is also important that decision makers understand the different uses and limitations of various methods of analyzing and interpreting the data, including models that attempt to forecast how the disease will continue to unfold.

The Coronavirus Aid, Relief, and Economic Security Act (CARES Act) includes a provision for us to, in general, conduct monitoring and oversight of the authorities and funding provided to address the COVID-19 pandemic and the effect of the pandemic on the health, economy, and public and private institutions of the United States.[2] This technology assessment discusses (1) collection methods and limitations of surveillance data for COVID-19 cases, hospitalizations, and deaths reported by CDC, (2) approaches for analyzing COVID-19 data, and (3) uses and limitations of forecast modeling for understanding COVID-19.[3]

To address collection of COVID-19 surveillance data, we obtained documentation on CDC's efforts to report data on cases, hospitalizations, and deaths from its website (including CDC's National Center for Health Statistics (NCHS) website), reports, and other publications. To address approaches for analyzing COVID-19 data, we reviewed available data and applied staff

[1]Social distancing measures are actions taken to decrease virus transmission by reducing contact among individuals within or between populations, such as by closing restaurants, businesses, and schools or by restricting travel and issuing shelter-in-place/stay-at-home orders. For more on this topic, see GAO, *Science & Tech Spotlight: Social Distancing During Pandemics*, GAO-20-545SP (Washington, D.C.: May 13, 2020).

[2]Coronavirus Aid, Relief, and Economic Security Act, Pub. L. No. 116-136, § 19010(b), 134 Stat. 281, 580 (2020).

[3]This report does not include other types of surveillance efforts conducted by CDC and some state and local health officials, such as COVID-19 outpatient and emergency department illness surveillance or recovered cases surveillance; however, we are examining these other types of data for our future work.

expertise in public health, epidemiology, and biostatistics. To address uses and limitations of forecast modeling for COVID-19, we used prior GAO work on disease modeling and specialist staff expertise to identify the purpose, structure, input data, and results of selected forecasting models that published technical documentation describing their methods and assumptions.

We conducted our work from May 2020 to July 2020 in accordance with all sections of GAO's Quality Assurance Framework that are relevant to technology assessments. The framework requires that we plan and perform the engagement to obtain sufficient, appropriate evidence to meet our stated objectives and to discuss any limitations to our work. We believe that the information and data obtained, and the analysis conducted, provide a reasonable basis for any findings and conclusions in this product.

1 Background

On March 11, 2020, the World Health Organization declared the COVID-19 outbreak a pandemic. As of June 18, 2020, the United States had approximately 2.1 million reported cases of COVID-19 and more than 100,000 reported deaths, according to federal agencies.[4] Parts of the nation have also experienced overwhelmed health care systems along with a rapid and severe deterioration in the economy. The United States has taken a number of actions to respond to the pandemic. These actions have spanned the whole of government as the pandemic has affected the economy, the educational system, and virtually every aspect of life.

There are several different types of coronaviruses, some of which are responsible for the common cold, and some of which cause severe respiratory illness and have high mortality rates.[5] In addition to COVID-19, other severe outbreaks of respiratory illness caused by coronaviruses in the past 20 years include Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome.

1.1 U.S. public health surveillance systems and COVID-19

Public health surveillance is the ongoing, systematic collection, analysis, and interpretation of health-related data essential to planning, implementing, and evaluating public health practice.[6] For surveillance of COVID-19, CDC—within the Department of Health and Human Services (HHS)—is using multiple systems run in collaboration with state, local, and academic partners to monitor COVID-19 in the United States.[7] For example, COVID-19 cases are tracked in the National Notifiable Diseases Surveillance System, which CDC uses to monitor about 120 diseases.[8] CDC and state health departments post the COVID-19 surveillance data to their websites, while other entities, such as academic institutions and media outlets, obtain these data from the states, CDC, and other sources and post them on their own COVID-19 websites. For example, USAFacts is a not-for-profit, nonpartisan civic initiative that aggregates data from CDC, state, and local public health agencies to publish time

[4]For more information, see https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html.

[5]For more information, see GAO, *Science & Tech Spotlight: Coronaviruses*, GAO-20-472SP (Washington, D.C.: Mar. 3, 2020).

[6]For more information, see https://www.cdc.gov/publichealth101/surveillance.html (accessed June 19, 2020).

[7]For more information, see https://www.cdc.gov/coronavirus/2019-ncov/covid-data/faq-surveillance.html, (accessed June 9, 2020)

[8]The National Notifiable Diseases Surveillance System provides the underlying data that public health officials at CDC need to monitor disease trends, study risk factors, evaluate prevention and control efforts, and target public health resources. Hospitals, laboratories, and health care providers send data to local and state public health departments, which then voluntarily submit data to CDC to include in the National Notifiable Diseases Surveillance System. For more information, see https://www.hhs.gov/cto/projects/national-notifiable-diseases-surveillance-system-modernization-initiative/index.html#:~:text=The%20National%20Notifiable%20Diseases%, accessed June 26, 2020.

series data on confirmed cases and deaths at the county level.

1.2 Infectious disease models

In general, a model uses equations and logic to simplify aspects of nature that can be complicated and difficult to understand. Models are generally not intended to be perfect representations of reality, but they can be used to test theories or predictions about how the real world works. Although the real world is complex and relevant information is not always known, models can help researchers and policymakers understand, compare, and predict how systems have behaved, are behaving, or may behave in the future.

Because models are simplifications of reality, model developers use assumptions to reduce the complexity of the systems the models represent or to fill in gaps in data or understanding. In developing a model, analysts write mathematical equations, logical rules, or both according to those assumptions to simplify the known reality of the phenomenon being modeled or to identify patterns from observed data. Analysts then feed data or estimates of key parameter values into the model as inputs. The model processes the inputs and delivers outputs in the form of point estimates, ranges, probabilities, graphs, tables, maps, or custom visualizations.

Infectious disease models are a subset of models that help researchers analyze the dynamics of a disease.[9] In the context of a public health response to a disease outbreak, models are used to understand the drivers of a particular outbreak, assess the risk of certain diseases, detect or forecast new outbreaks, and investigate the potential effects of public health interventions. They are also used when planning for future infectious disease outbreaks, such as in development of vaccines or other medical countermeasures, scenarios for responder exercises, or optimization of medical supplies. Additionally, models are used for academic purposes to more fully understand diseases or public health tools, but not necessarily in the context of a public health response to a disease outbreak or explicit planning for a future response by public health professionals. They have played a prominent role in the COVID-19 response in the United States and abroad—for example, in projecting new infections, deaths, and the need for health care resources.[10]

[9]For more information, see GAO, *Infectious Disease Modeling: Opportunities to Improve Coordination and Ensure Reproducibility*, GAO-20-372 (Washington, D.C.: May 13, 2020).

[10]For more information, see GAO, *Science & Tech Spotlight: COVID-19 Modeling*, GAO-20-582SP (Washington, D.C.: June 4, 2020).

2 Public Health Surveillance Data on COVID-19 That CDC Collects Have Important Limitations

Much uncertainty exists in the COVID-19 surveillance data that public health and other officials, academia, media outlets, and others are reporting—in some cases for policy and decision-making purposes. In particular, visual representations of the data can be misleading without appropriate context in terms of data completeness and reliability. To properly interpret the data, it is important to consider these limitations and the context of the different types of data being presented and analyzed in various reports and media coverage. This section of our report describes collection methods and limitations related to CDC's COVID-19 surveillance data, including reported cases, hospitalizations, and deaths.

2.1 Case data

2.1.1 Data collection

In general in the United States, health care providers, hospitals, laboratories, and others report cases to local public health authorities. Local health departments report cases to their state or territory health department, which reports them to CDC. The surveillance system relies on collaboration and participation between CDC, states, health care providers, laboratories, and others. Requirements as to what data are to be reported and by whom (e.g., which diseases or test results must be reported) are established at the state, territory, and local levels, but notification to CDC is voluntary. In addition to the laboratory test results, CDC and state and local health officials collect several types of information about the cases, including demographic (e.g., age), clinical (e.g., symptoms), and exposure (e.g., contact with a known infected person) information.

On April 5, 2020, the Council of State and Territorial Epidemiologists published an interim position statement to create, along with CDC, a standardized case definition for COVID-19 and to add COVID-19 to the list of National Notifiable Conditions in CDC's National Notifiable Diseases Surveillance System.[11] This case definition specifies that cases may be classified as confirmed or probable. Confirmed cases are those that are diagnosed through diagnostic laboratory viral testing for the SARS-CoV-2 virus.[12] Probable cases do not have a confirmed viral test, but

[11]Council of State and Territorial Epidemiologists, *Standardized Surveillance Case Definition and National Notification for 2019 Novel Coronavirus Disease (COVID-19)*, available at https://www.cste.org/news/500750/CSTE-Interim-Position-Statement-COVID-19-Case-Definition-and-Addition-to-the-NNC-List.htm, accessed June 30, 2020.

The Nationally Notifiable Diseases Surveillance System is a passive surveillance system that relies on voluntarily reporting and the uncompensated initiative of reporting sources. The incompleteness of data is a recognized limitation that CDC told us is accepted in exchange for the reduced cost of conducting passive surveillance. Using these data, numerous reports characterizing the epidemiology of COVID-19 have been published in CDC's Morbidity and Mortality Report (www.cdc.gov/mmwr). Active surveillance is discussed in section 2.2.

[12]There are generally two types of tests available for COVID-19—(1) viral tests for diagnosing current infection, and (2) antibody tests for past infection. For more information, see GAO, Science & Tech Spotlight: COVID-19 Testing, GAO-20-584SP (Washington, D.C.: May 20, 2020).

meet other case definition criteria.[13] CDC also developed a national case report form to standardize how providers and public health officials collect demographic, clinical, and epidemiologic information.[14] According to CDC, the case report form was revised in May 2020 to include the new case definition and update data collection elements to reflect changing U.S. COVID-19 epidemiology, including additional questions on health care personnel.

2.1.2 Limitations and considerations for analysis and interpretation

In June 2020, we reported that CDC's COVID-19 testing data were not complete or consistent. Specifically, delays in testing during the early pandemic stages resulted in limited information on the spread of COVID-19 in communities, and the sources of testing data CDC had used changed with changes in testing practices over time.[15] We discuss these and other limitations as they relate to COVID-19 case surveillance in the sections below.[16]

Consistency. Many factors affect the consistency of case data across time, location, and subpopulations, including variations in availability of tests and related supplies; health care providers' decisions to test and their testing capacity; criteria for testing;[17] whether an infected person seeks care and is tested; and capacity for case contact tracing.[18] In addition, jurisdictions may vary in the extent to which they report probable cases; cases that meet presumptive laboratory testing requirements may be considered probable if they meet other clinical and epidemiological evidence. Since April 14, 2020, CDC's case reporting has included both confirmed and probable cases. According to CDC, case surveillance counts would increase slightly with the inclusion of probable cases that were not reported until the new Council of State and Territorial Epidemiologists case definition, which CDC addressed by displaying confirmed and probable cases separately on its website as of June 2, 2020.

Completeness. Case surveillance data are incomplete for a number of reasons and are

[13]A probable case (1) meets clinical criteria and epidemiologic evidence with no confirmatory laboratory testing performed for COVID-19, (2) meets presumptive laboratory evidence (i.e., antibody test) and either clinical criteria or epidemiologic evidence, or (3) meets vital records criteria with no confirmatory laboratory testing performed for COVID-19. For details, see Centers for Disease Control and Prevention, *Coronavirus Disease 2019 (COVID-19) 2020 Interim Case Definition, Approved April 5, 2020*, accessed June 23, 2020, https://wwwn.cdc.gov/nndss/conditions/coronavirus-disease-2019-covid-19/case-definition/2020/ .

[14]This form is available at https://www.cdc.gov/coronavirus/2019-ncov/downloads/pui-form.pdf, accessed April 21, 2020.

[15]GAO, *COVID-19: Opportunities to Improve Federal Response and Recovery Efforts,* GAO-20-625 (Washington, D.C.: June 25, 2020).

[16]CDC's testing tracking provides data on the total number of tests reported and their results (regardless of how many times each person is tested), while case surveillance tracks people who are diagnosed with COVID-19.

[17]Testing availability was particularly problematic earlier in the pandemic (see GAO-20-625). CDC's guidelines for testing expanded over time, but until more recently, testing still largely remained limited to symptomatic individuals, health care workers, and others at high risk for exposure, due to ongoing limited testing capacity.

[18]In case investigation, public health staff work with a patient to help them recall everyone with whom they have had close contact during the time frame while they may have been infectious. Public health staff then begin contact tracing by warning these exposed individuals (contacts) of their potential exposure as rapidly and sensitively as possible. Contact tracing also often involves testing of contacts who are known to have been exposed and may or may not have symptoms.

an undercount of the true number of cases, according to CDC and others. For example, infected people with symptoms may not get tested due to limited testing availability and access, or they may not seek care to be tested. Current evidence also suggests that many infected people may be asymptomatic or have only mild symptoms and not seek health care and get tested, which affects the ability of case surveillance to identify all infections.

Completeness of case surveillance data also depends on the availability of tests, as well as the availability of staff and other resources needed to conduct and report the test results. Completeness of reporting also depends on whether health care providers and public health agencies are able to keep up with the administrative burden of reporting cases, according to CDC. COVID-19 viral testing in the United States was initially developed by CDC and conducted at some U.S. public health laboratories. The Food and Drug Administration issued a new policy on February 29, 2020, which it subsequently modified, to help expedite availability of viral testing for COVID-19.[19] These policy changes, along with increased testing sites, staffing, and other resources, have led to expanded viral testing capacity. For further discussion of issues related to viral testing, see our June 2020 report.[20]

Completeness of case data for more recent time periods also depends on the amount of time it takes to obtain test results. The amount of time required between specimen collection and obtaining test results varies by type of test as well as the laboratory performing the test, with some tests taking several days for results to return, while newer more rapid diagnostic tests at the point-of-care (such as a doctor's office) can provide results in hours instead of days. Additional delays in reporting can occur as data go from the local health departments to the states to CDC. In addition, CDC posts official case counts after a confirmation process with each jurisdiction. This process may result in discrepancies between CDC's case counts and other sources, including counts posted on state health department websites.

Further, demographic and other epidemiological information on cases is limited and varies over time and by population and geographic location. For example, as of July 21, 2020, CDC reported that of 3,819,139 cases, data were collected from 2,451,286 people, of which race/ethnicity was only available for 1,377,305 people (36 percent). CDC noted that it is working with states to provide more information on race/ethnicity for reported cases. On June 4, 2020, HHS—under authority granted to it in the CARES Act—announced new guidance that specifies that additional data, including race, ethnicity, age, and sex, must be reported to HHS by laboratories when they submit COVID-19 test results beginning August 1, 2020.[21] According to HHS,

[19]Food and Drug Administration, *Policy for Diagnostic Tests for Coronavirus Disease-2019 during the Public Health Emergency Immediately in Effect Guidance for Clinical Laboratories, Commercial Manufacturers, and Food and Drug Administration Staff*, March 16, 2020, which superseded *Policy for Diagnostics Testing in Laboratories Certified to Perform High-Complexity Testing under Clinical Laboratory Improvement (CLIA) prior to Emergency Use Authorization for Coronavirus Disease-2019 during the Public Health Emergency*, issued February 29, 2020.

[20]GAO-20-625.

[21]CARES Act, Pub. L. No. 116-136, div. B, tit. VIII, § 18115(a), 134 Stat. at 574; Health and Human Services, *COVID-19*

the new requirements will enable HHS to ensure that all groups have equitable access to testing and to accurately determine the burden of infection on vulnerable groups. To the extent that these additional data are incorporated into laboratory information systems and case reporting systems, the changes could also improve completeness of demographic information for cases.

The above issues concerning consistency and completeness of case data complicate comparisons of COVID-19 case counts and rates among geographic areas and populations, and may also make assessing county-, state- and national-level trends over time difficult. Some counties have also been testing and tracking COVID-19 cases for a longer period of time than others. For counties that began tracking cases later, it is difficult to isolate the actual time period in which cases began to increase because the increase may have started before tracking began.[22] Testing for COVID-19 is necessary for knowing the full extent of the disease in the population. A low rate of testing in an area could mean that additional cases could exist undetected in a community. Variations in case identification, data collection, and reporting processes may play a role in observed differences in case counts between geographic areas and subpopulations and over time. COVID-19 data issues can also affect researchers' ability to conduct reliable analyses, such as analysis of the effects of social distancing policies on the number of COVID-19 cases over time.

Similarly, inconsistent and incomplete information on cases, including demographic and other characteristics such as other health conditions, can affect researchers' ability to fully ascertain the extent to which different population subgroups may be disproportionately affected by COVID-19. When analyzing COVID-19 data, it is important to consider rerunning analyses as the data change, which can lead to changes in results depending on the extent of such revisions.

2.2 Hospitalization data

2.2.1 Data collection

CDC is conducting surveillance for laboratory-confirmed hospitalizations through two existing programs in 14 states, according to its website. CDC monitors laboratory-confirmed COVID-19-associated hospitalization rates through the COVID-19-Associated Hospitalization Surveillance Network (COVID-NET).[23] COVID-NET conducts all-age, population-based surveillance for laboratory-confirmed COVID-19-associated hospitalizations in more than 250 acute care hospitals in 99 counties in the 10 Emerging Infections Program states and four Influenza Hospitalization Surveillance Project states.[24]

Pandemic Response, Laboratory Data Reporting: CARES Act Section 18115 (June 4, 2020).

[22]CDC provides county-level data courtesy of USAFacts.org.

[23]COVID-NET is an active surveillance system. The system is a component of a larger Emerging Infections Program, which is funded by CDC to conduct population-based surveillance in geographically defined catchment areas where near complete case ascertainment and data capture allow for accurate estimation of incidence rates in various populations. CDC told us that COVID-NET and the Emerging Infections Program are more expensive than the passive case surveillance described in section 2.1, but yield high quality data.

[24]The 10 Emerging Infections Program states are California, Colorado, Connecticut, Georgia, Maryland, Minnesota, New Mexico, New York, Oregon, and Tennessee. The four Influenza Hospitalization Surveillance Project states are Iowa, Michigan, Ohio, and Utah. The project began during the 2009–2010

In total, COVID-NET covers approximately 10 percent of the U.S. population, according to CDC's website. CDC's website also notes that the counties covered in COVID-NET are located in all 10 HHS regions, and that the designated COVID-NET surveillance area is generally similar to the U.S. population by demographics; however, the information might not be generalizable to the entire country.[25]

To be counted as a COVID-19 case, a patient must be a resident of a designated catchment area of 99 counties in 14 states and hospitalized within 14 days of a positive SARS-CoV-2 test.[26] Testing is performed at the discretion of health care providers. Hospital and COVID-NET staff identify hospitalizations through active review of clinical laboratory databases and hospital admission and infection control practitioner records. CDC uses the actively-collected data to estimate demographic-specific hospitalization rates on a weekly basis and to describe characteristics of persons hospitalized with COVID-19 illness. Hospital and COVID-NET staff also review patient charts to determine if certain categories of underlying medical conditions are recorded in the chart at the time of hospitalization. In addition to the 14 states participating in CDC's COVID-19 hospitalization surveillance, other efforts are underway to track hospitalizations and other hospital-based indicators in states and localities.[27]

2.2.2 Limitations and considerations for analysis and interpretation

Consistency. Health care providers and facilities within the surveillance catchment area of 14 states may vary in their testing practices, and testing availability may also vary across and within the 14 participating sites.

Completeness. Similarly, provider testing practices and testing availability and capacity can affect the completeness of COVID-19 hospitalization data. CDC's website notes that undercounting of cases in COVID-NET is likely. Availability of demographic and epidemiological information is also limited in the surveillance system. For example, CDC's COVIDView Weekly Surveillance Summary, updated July 17, 2020, reported a total of 37,052 laboratory-confirmed COVID-19-associated hospitalizations between March 1, 2020, and July 11, 2020. However, 94 percent (34,669) included information on

influenza season to enhance surveillance during the 2009 H1N1 pandemic, and different states participate in different years of enhanced surveillance.

[25]The HHS Office of Intergovernmental and External Affairs hosts 10 regional offices that directly serve state and local organizations.

[26]A catchment area is a defined geographic area served by an institution. In the case of COVID-NET, it refers to the states that participate in these hospital surveillance programs

[27]For example, see University of Minnesota "*COVID-19 Hospitalization Tracking Project*," accessed 6/30/2020, https://carlsonschool.umn.edu/mili-misrc-covid19-tracking-project, a project that was launched on March 26, 2020 to track and report daily hospitalizations from all 50 states. As another example, CDC's National Healthcare Safety Network (NHSN) added a COVID-19 module for tracking hospital-based measures. CDC told us that hospitals reported key indicators of hospital capacity to NHSN, including available hospital beds, ICU beds, and ventilators—and the percentage of COVID-19 patients using these resources. Hospitals also reported staffing and PPE supply shortages to NHSN. By early May 2020, more than 3,500 hospitals were reporting these data daily to CDC's NHSN, representing approximately 60% of all hospitals in the country, according to CDC. However, CDC stated that effective July 15, 2020, hospitals no longer report COVID-19 capacity, staffing, and supply-related data to CDC's NHSN, and that hospital data can be accessed from HHS Protect, a unified data set maintained by HHS

race/ethnicity, while only 26 percent (9,736) included information on underlying medical conditions such as hypertension and obesity.

In addition, results for COVID-NET may not be representative of populations outside the surveillance catchment area. A comparison of characteristics of populations within and outside the catchment area may be helpful for understanding whether results may be applicable to the noncatchment areas.

The COVID-NET data limitations noted above can affect analysts' ability to interpret trends over time and across subpopulations. However, according to CDC, all identified COVID-19 hospitalizations within the COVID-NET catchment area will ultimately have complete medical chart abstractions.

2.3 Mortality data

2.3.1 Data collection

Data on COVID-19 deaths in the United States are collected from two sources: 1) follow-up on cases reported through the case surveillance system, and 2) National Center for Health Statistics (NCHS) reporting based on death certificates, which serve as the basis for mortality statistics from all causes of death.[28]

1. The COVID-19 death count shown on CDC's *Cases in the U.S*. web page includes preliminary deaths reported daily by state, local, and territorial health departments through the COVID-19 case surveillance system.[29] This count reflects the most up-to-date information received by CDC based on preliminary reporting from health departments.
2. Current and prior weekly provisional COVID-19 death counts that CDC reports from NCHS are updated Monday through Friday with information collected from death certificates. These data are based on official death certificates and represent the most accurate death data, according to CDC. Certifiers (e.g., attending physicians, medical examiners, and coroners) complete the death certificate and generally file it with local and state registrars, sometimes simultaneously, and then registrars submit it to NCHS.[30]

As with COVID-19 cases, COVID-19 deaths can be classified as confirmed or probable when they are reported through the case surveillance system. For NCHS COVID-19 deaths reported on death certificates, certifiers are asked to use their best clinical judgement about the cause or causes of death using all available evidence, including a laboratory test if available.

[28]According to CDC's website, COVID-19 deaths are identified using the ICD–10 code U07.1. Deaths are coded to U07.1 when COVID-19 is reported as a cause that contributed to death on the death certificate.

[29]See Centers for Disease Control and Prevention, *Cases in the U.S.*, accessed June 23, 2020, https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html.

[30]According to CDC's website, the mission of NCHS—the nation's principal health statistics agency—is to provide statistical information, including mortality data, which will guide actions and policies to improve the health of the American people.

2.3.2 Limitations and considerations for analysis and interpretation

Consistency. Several factors affect the consistency of data on COVID-19 deaths. State and local entities may differ in the extent to which they count probable COVID-19 deaths in addition to laboratory-confirmed COVID-19 deaths in the death counts reported through the case surveillance system. States that had not reported probable COVID-19 deaths from the case surveillance system prior to April 14, 2020, would have seen an uptick in deaths from this system if they began to include probable deaths after this time. In addition, because COVID-19 is a respiratory disease, COVID-19 deaths may be misclassified as pneumonia or influenza in the absence of positive test results. There may also be variations in death certificate certifiers in classifying COVID-19 as a cause of death on the death certificate, because death certifiers make cause-of-death determinations based on their best clinical judgement that is informed by their experience and knowledge.

Completeness. COVID-19 mortality data are incomplete in at least two ways. First, the data can vary for a particular day because states report death counts at different time intervals. For NCHS deaths, NCHS continuously revises provisional death counts as it receives new and updated death certificate data from the states. NCHS reports that 63 percent of U.S. deaths are reported within 10 days of the date of death, but there is variation within states. Twenty states report over 75 percent of deaths within the first 10 days, while three states report less than 1 percent of deaths within 10 days.[31] NCHS updates the provisional counts of total COVID-19 deaths and deaths by state daily and updates the provisional counts of COVID-19 deaths tabulated by demographic and other geographic characteristics every week. Reporting on provisional COVID-19 mortality data from NCHS currently lags by an average of 1–2 weeks with a range of 1–8 weeks.[32] Therefore, the provisional death counts may not include all deaths that occurred during a given period, especially for more recent periods. Death counts from earlier weeks are continually revised and may increase or decrease as new and updated death certificate data are received. For these reasons, provisional COVID-19 death counts may differ from those in other published sources, such as media reports or the *Cases in the U.S.* web page. States may also differ in how quickly they receive reports of COVID-19 deaths from local jurisdictions and report them to CDC as part of the case surveillance system. CDC's case surveillance provides updated daily death totals on its website, and reporting delays depend on how quickly states report their counts to CDC. As a result of these differences in the timing of reporting through the case surveillance system and to NCHS, comparisons of death counts among states can be problematic.

Second, NCHS provides the death counts by selected demographic and geographic characteristics on a weekly basis, including age, sex, race, and Hispanic origin. It also

[31]See Centers for Disease Control and Prevention, *Technical Notes: Provisional Death Counts for Coronavirus Disease (COVID-19)*, accessed June 29, 2020, https://www.cdc.gov/nchs/nvss/vsrr/covid19/tech_notes.htm.

[32]Final death counts are published approximately 11–12 months from the end of the data year.

includes place of death (e.g., hospital, home, or nursing home) and comorbid conditions that contributed to death. However, there is variation in whether jurisdictions provide this information on deaths reported through the case surveillance system. Among those that do, the data are not always complete. These data are needed to understand whether certain groups are experiencing higher-than-expected death rates from COVID-19. CDC attempted to rectify this gap by issuing the June 4 guidance, discussed earlier, requiring all states to report demographic data with test data beginning August 1, 2020. To the extent that these additional data are incorporated into the laboratory information systems and the death reporting systems, the changes could also improve completeness of demographic information for death data.

The extent of any net undercounting or overcounting of COVID-19 deaths is unknown. Undercounting could result from a lack of testing to confirm COVID-19 deaths, while overcounting could result from deaths from other causes being classified as COVID-19 deaths. The availability of testing could have a notable effect on death counts reported through the case surveillance system prior to the inclusion of probable cases in the case definition.

The accuracy and completeness of COVID-19 case and death counts affect the case fatality ratio—that is, the ratio of total deaths to total cases. This ratio is a way of measuring the proportion of people who become infected that will die from the disease, and it affects disease modeling and forecasting (discussed in chapter 4). If the number of actual cases is higher than the number of reported cases, the fatality ratio may be lower—that is, better—than analysts have estimated. In contrast, if COVID-19 deaths are undercounted, the fatality ratio may be higher—that is, worse—than the estimation.[33]

[33]For more information, see Anthony S. Fauci, H. Clifford Lane, and Robert R. Redfield, "COVID-19—Navigating the Uncharted," *New England Journal of Medicine*, vol. 382, no. 13 (Mar. 26, 2020): 1268–1269.

3 Different Analytical Approaches Yield Different Insights

COVID-19 data can be presented and analyzed in different ways, and these approaches often yield different insights. This report section highlights the uses of and potential sources of difference among some of these approaches. These analytical approaches include using data on cases, hospitalizations, and deaths; aggregating data; incorporating demographics into analyses to compare across population groups; using approaches that address the imperfect reporting of COVID-19 deaths; and comparing COVID-19 deaths and deaths from other causes.[34]

Table 1: Examples of uses for data on cases, hospitalizations, and deaths

Measures	This measure may be useful when
Cases	Examining the total spread of the virus
Hospitalizations	Examining health care system capacity and the severity of COVID-19
Deaths	Examining the severity of COVID-19

Source: GAO. | GAO-20-635SP

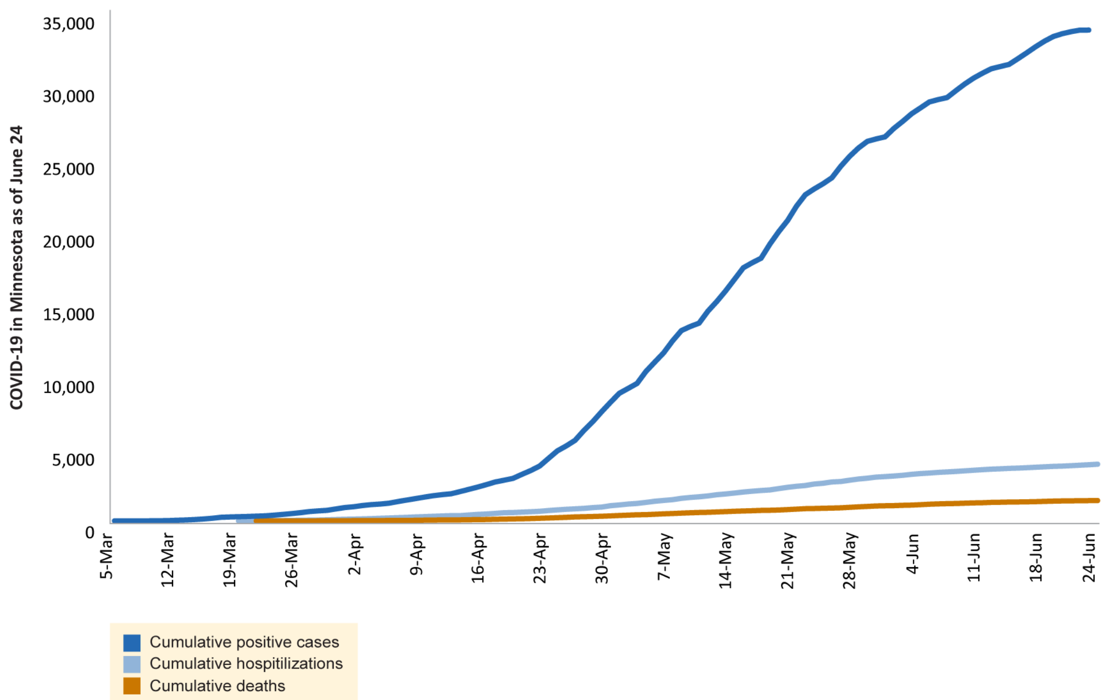
3.1 Measures of cases, hospitalizations, and deaths serve different purposes

COVID-19 measures of cases, hospitalizations, and deaths help analysts to understand different aspects of COVID-19 (see table 1). COVID-19 cases reflect the spread of the virus, while hospitalizations show the effect on the health care system and deaths show the severity of COVID-19. (We discuss how these measures are collected and their limitations in chapter 2). While all of these measures are useful, they often show different trends, and the decision of which measure to use affects the conclusions that can be drawn from the data. Understanding the potential uses of each of the measures allows analysts to choose the one that is best suited to answer the question they are addressing.

Trends for cases, hospitalizations, and deaths may be different for different geographic areas. Cases have the broadest scope because they encompass the other measures (e.g., a hospitalization is also a case) and will be the largest count compared to hospitalizations and deaths in any particular area, but each measure may have a different rate of increase or decrease. For example, figure 1 shows the cumulative numbers of reported positive cases, hospitalizations, and deaths in Minnesota from March 5 to June 24, as reported by June 24. The numbers of deaths and hospitalizations are a small fraction of the overall number of cases, and the increase in the number of cases is steeper than the other two measures. Although the cumulative numbers of positive cases, hospitalizations, and deaths have continued to increase since June 24, 2020, this pattern continues.

[34]This section uses data reported by CDC on COVID-19 cases. For analyses of race/ethnicity, we used data from the New York City Department of Health and Mental Hygiene, which was a consistent and reliable source of data on the race/ethnicity of deaths at the time of our analysis. We used NCHS data for an illustrative example of excess deaths. For an example of cases, hospitalizations, and deaths, we used data from the Minnesota Department of Health.

Figure 1: Cumulative reported COVID-19 cases, hospitalizations, and deaths in Minnesota as of June 24, 2020



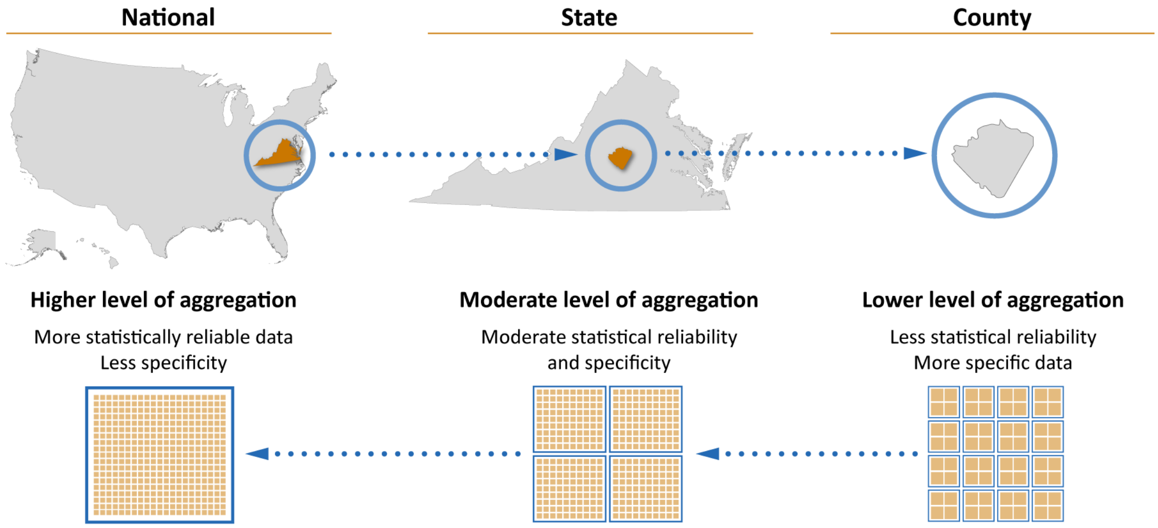
Source: GAO analysis of the Minnesota State Department of Health data. | GAO-20-635SP

3.2 Aggregating data can improve reliability but obscure trends

One way to understand the magnitude of COVID-19 cases, hospitalizations, and deaths is to view the data at different levels of aggregation. Specifically, national estimates may be disaggregated to view regional, state, and county estimates, and vice versa (see fig. 2).

While higher levels of aggregation provide an overall sense of magnitude, this approach may obscure important trends, as some areas of the country have been affected more than others. Examining data at a lower level of aggregation, such as at the county level, allows analysts to see how different geographic areas are affected.

Figure 2: Data can be aggregated at different levels of geographic specificity

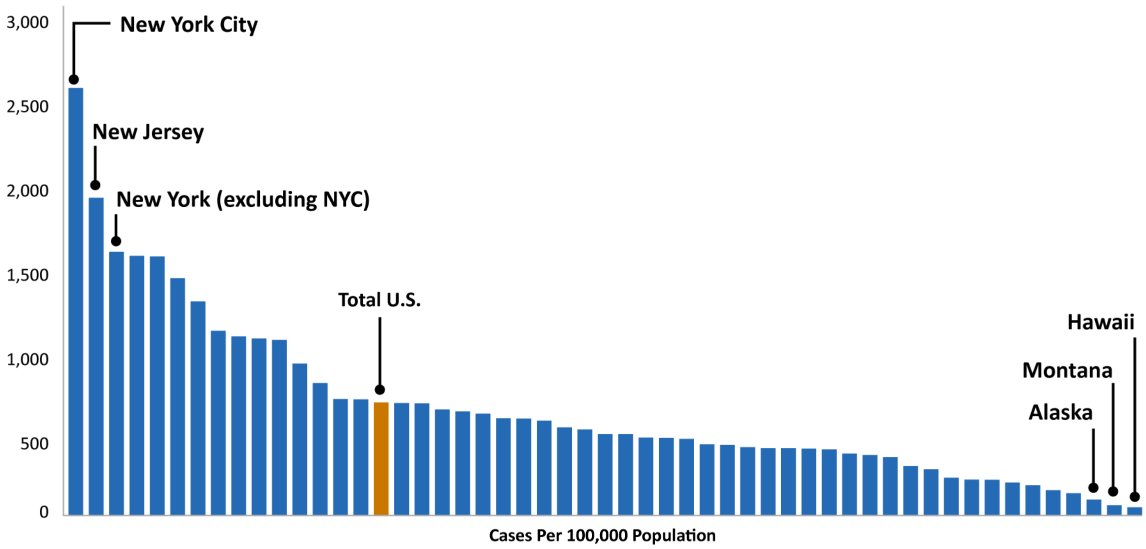


Source: GAO. | GAO-20-635SP

For example, according to CDC data as of June 19 2020, while the national case rate was about 673 reported COVID-19 cases per 100,000 people, the rate ranged from nearly 1,900 in New Jersey to about 50 in Hawaii (see fig. 3). There was wide variation in the case rate by state that was hidden in the national estimate. Although case rates in many of the jurisdictions have changed substantially since June 19, 2020, there continue to be large differences in the case rates across states. Additionally, because reporting decisions and testing capacity vary across states, aggregating across states would mean combining inconsistent data and may not be appropriate.

However, one drawback to analyzing data for smaller areas is that data based on fewer people may be less statistically reliable, resulting in more unstable trends. Unstable trends are those for which random or small movements in the numbers make it difficult to see any trends. Using a 1-week moving average instead of daily numbers can help to even out these random spikes in the data. While national data based on a greater number of people may also be subject to unstable trends, this instability could be resolved by using a moving average to even out any spikes in the data, rather than because the data are based on too few people, which cannot simply be addressed by using a moving average. There are various techniques for addressing unreliability in estimates from low population areas. The most straightforward is to aggregate the data to a larger geographic area. For example, less reliable county-level data may be aggregated to produce more reliable state estimates. A more complex method involves small area estimation, in which county-level estimates use information from other counties to generate more stable estimates. Areas of intense transmission may have enough cases, hospitalizations, and deaths to improve reliability, even if they come from smaller areas.

Figure 3: Reported COVID-19 cases per 100,000 population, as of June 19, 2020



Source: GAO analysis of data reported by the Centers for Disease Control and Prevention. | GAO-20-635SP

Additional data limitations relating to timeliness and local-level reporting also affect analysis of publicly available county-level data on COVID-19 cases and deaths. For example, official time series data at the county level are not available, and CDC does not have its own data but rather directs data users to a third party, USAFacts, for these data on cases and deaths.[35] States and counties may also vary in how they assign a case to a given county for their own jurisdictional reporting, with some assigning cases to the patient's county of residence and others to the county in which a patient is treated for COVID-19. In addition, the department can classify the patient into a separate "unknown" county designation when the county of residence is unknown or is pending determination.[36] As a result, the measurement of the county of cases and deaths could cause misclassification when examining the data by county.

State-level data provide a good balance between reliability and meaningful insight. In general, states are large enough to address some of the reliability problems associated with smaller areas but small enough to capture regional variation and inform policies for social distancing and reopening adopted by state governments. Further, state-level analysis avoids the problem of aggregating data on COVID-19 cases and deaths that states are collecting in different ways, which may not allow for meaningful aggregation to regional or national levels. However, state-level data may not be appropriate for all

[35]Time series data—a sequence of measurements taken at regular points in time—are fundamental to understanding the underlying structure of observed processes and performing both descriptive and explanatory analyses, as well as intervention analysis. USAFacts is a not-for-profit nonpartisan civic initiative which aggregates data from CDC, state, and local public health agencies to publish time series data on confirmed cases and deaths at the county level.

[36]For standardized case and death reporting to CDC and NCHS, the county of residence and county where the case was identified or death occurred are both reported.

purposes, such as when cities or counties need local data to support policy decisions. Finally, national estimates may be more appropriate for making high-level comparisons, such as those among countries, or for developing a single measure to summarize conditions in the whole country.

3.3 Incorporating demographics into analyses allows comparison across population groups

Similar to disaggregating analyses by geographic area, disaggregating analyses by demographic group can shed light on groups that may be disproportionally affected by COVID-19. COVID-19 does not affect all population groups in the same way. For example, the disease may affect some groups differently depending on their race, age, and underlying health.[37] Further, there are a number of ways of adjusting COVID-19 measures that allow analysts to make meaningful comparisons across population groups. Specifically, data can be provided as basic counts or numbers, but they can also be presented in terms of proportions, unadjusted rates, and adjusted rates (see table 2).

However, in order to analyze data in these ways, analysts must have sufficient quality information on demographic characteristics. CDC's COVID-19 case report form on its website as of June 22, 2020, collects information on demographics, so that CDC

Table 2: Comparison of common COVID-19 measures

Measure	Best uses	Limitations
Basic counts or numbers	Understanding the overall magnitude of cases, hospitalizations, and deaths	Does not account for population size or demographics
Proportions	Making comparisons across population groups of different sizes when there is a substantial amount of missing demographic data	Does not account for the different age distributions of the populations being compared (among other characteristics)
Unadjusted rates	Making comparisons across population groups of different sizes	Does not account for the different age distributions of the populations being compared (among other characteristics)
Age-adjusted rates	Making comparisons across population groups of different sizes and different demographic characteristics	Not as easy to interpret Requires measurement of age Does not adjust for demographics other than age

Source: GAO. | GAO-20-63SP

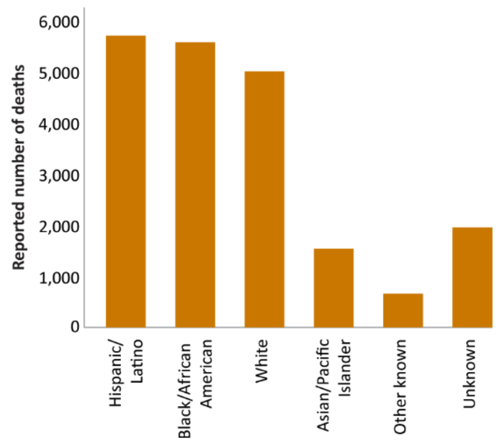
[37]A CDC report shows that COVID-19 hospitalizations and deaths have affected people in different demographic groups in different ways. See Jeremy A.W. Gold et al., "Characteristics and Clinical Outcomes of Adult Patients Hospitalized with COVID-19—Georgia, March 2020," *Morbidity and Mortality Weekly Report,* Centers for Disease Control and Prevention (May 8, 2020): 545–550.

and state health departments can examine the distribution of cases and deaths by these demographic characteristics.[38] Granular details of these data are often not released to the public due to concerns about statistical reliability and confidentiality, so the public relies on these organizations to produce meaningful estimates that help to make these comparisons. Missing demographic data make it challenging to disaggregate or make the adjustments that facilitate comparisons across groups. As discussed previously, HHS announced that as of August 1, 2020, all laboratories conducting tests to diagnose COVID-19 will be required to include demographic data such as race, ethnicity, age, and sex.[39] To the extent that these additional data are incorporated into laboratory information systems and case reporting systems, the changes could improve completeness of demographic information for cases. In the following subsections, we provide examples of different approaches to incorporate demographics into measures of COVID-19 deaths.

3.3.1 Counts or numbers of deaths

Counts are simply the number of deaths for each population group. For instance, total COVID-19 deaths in New York City can be disaggregated by race/ethnicity. Of the over 20,000 reported people who had died of confirmed or probable COVID-19 in New York City as of May 13, 2020, 5,600 were Black/African-American, 5,700 were Hispanic/Latino, and 5,000 were White, based on data from the New York City Department of Health and Mental Hygiene (see fig. 4). However, counts can be misleading when making comparisons among population groups of different sizes. For instance, the number of deaths in the Black/African-American and Hispanic/Latino populations in New York City were similar. These counts give the false impression that these populations have been equally affected by COVID-19 deaths because counts do not consider population size. The Hispanic/Latino population in New York City is larger than the Black/African-American population, around 2.5 million versus 1.9 million, respectively. Given the differences in population size, we would expect a greater number of deaths from the Hispanic/Latino population if both groups were equally affected.

Figure 4: Number of reported COVID-19 deaths by race/ethnicity in New York City, as of May 13, 2020



Source: GAO analysis of New York City Department of Health and Mental Hygiene data. | GAO-20-635SP

Note: Hispanic/Latino includes people of any race. "Other known" includes data on persons who identify as Native American/Alaska Native, or two or more races.

[38]Centers for Disease Control and Prevention, *Information for Health Departments on Reporting Cases of COVID-19*, accessed June 22, 2020, https://www.cdc.gov/coronavirus/2019-ncov/php/reporting-pui.html.

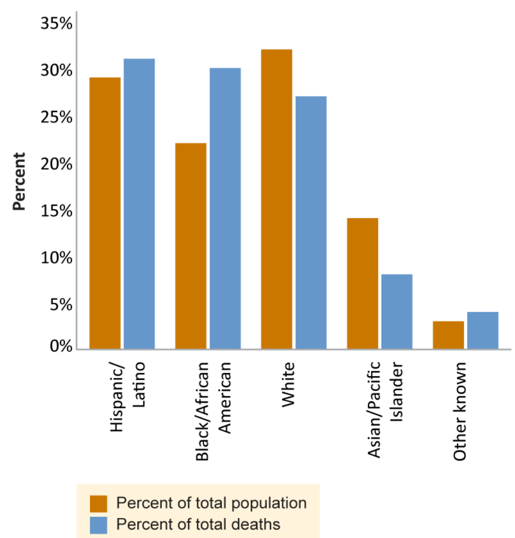
[39]Department of Health and Human Services, *COVID-19 Pandemic Response, Laboratory Data Reporting: CARES Act Section 18115*, (June 4, 2020), accessed June 26, 2020, at https://www.hhs.gov/sites/default/files/covid-19-laboratory-data-reporting-guidance.pdf.

It is important to note in figure 4 above that there was a substantial number of deaths with unknown race/ethnicity—nearly 2,000, or 10 percent. These 2,000 deaths were not counted under their actual race/ethnicity category, which means the counts for all known race and ethnic groups are underestimates. Further, some race and ethnic groups are generally more likely to be classified as unknown race than others. For example, an NCHS study of the accuracy of race/ethnicity reporting on death certificates found that accuracy was high for the White and Black/African-American populations.[40] However, the study also found that the misclassification was high for Native American/Alaska Native populations and moderate for the Hispanic/Latino and Asian/Pacific Islander populations. These results suggest that the nearly 2,000 deaths with unknown race/ethnicity may be less likely to be Black/African American or White than the other race/ethnicity categories.

3.3.2 Proportions of deaths

Unlike basic counts, calculating the proportion of deaths represented by different demographic groups allows analysts to see whether COVID-19 is affecting some groups disproportionately and partially accounts for deaths with unknown race/ethnicity. These percentages are calculated as the number of COVID-19 deaths for a demographic group in a particular geographic area divided by the total number of COVID-19 deaths in that area. These percentages must be compared side-by-side with the distribution of the population by racial/ethnic groups, which provides the expected percentage of deaths due to COVID-19 if all groups were affected equally. Groups that have a higher percentage of deaths relative to their percentage of the population were affected disproportionately more. For example, individuals identified as Black/African-American made up about 22 percent of the population in New York City from 2014 through 2018, but the Black/African-American demographic represented about 27 percent of COVID-19 deaths as of May 13, 2020, according to our analysis of New York City Department of Health and Mental Hygiene data (see fig. 5).

Figure 5: Distribution of the population and reported COVID-19 deaths in New York City, as of May 13, 2020



Source: GAO analysis of New York City Department of Health and Mental Hygiene and New York City Population FactFinder data. | GAO-20-635SP

Note: Deaths were as of May 13, 2020. Population estimates obtained from the New York City Population FactFinder were from the 2014–2018 American Community Survey. Hispanic/Latino includes people of any race. "Other known" includes data on persons who identify as Native American/Alaska Native, or two or more races.

[40]Elizabeth Arias et al., *The Validity of Race and Hispanic-Origin Reporting on Death Certificates in the United States: An Update*, Vital and Health Statistics Series 2, no. 172 (Hyattsville, Md.: National Center for Health Statistics, August 2016).

This method allows analysts to exclude unknown deaths from their analysis by focusing on the number of deaths within particular racial/ethnic groups in relation to the proportion of the population those groups represent. However, calculating the proportion of deaths only partially accounts for the number of deaths with unknown race/ethnicity. By limiting the analysis to those with known race/ethnicity, in effect we assume that deaths with unknown race/ethnicity are proportionally distributed across all known race/ethnicity groups. However, as previously discussed, people of unknown race/ethnicity are generally less likely to be Black/African-American or White than the other race/ethnicity categories. Therefore, calculating the percentage of deaths only partially accounts for the missing demographic data.

3.3.3 Unadjusted death rates

Similar to calculating proportions of deaths by demographic group, calculating death rates (also known as unadjusted death rates or deaths "per capita") allows analysts to compare population groups of different sizes. These death rates are calculated as the number of deaths in a given population group divided by the group's population size. The rates are usually multiplied by 100,000 to obtain deaths per 100,000 population. Based on data from the New York City Department of Health and Mental Hygiene, the COVID-19 death rate in New York City was approximately 240 deaths per 100,000 population as of May 13, 2020—calculated as 20,000 COVID-19 deaths divided by the population of New York City (8,340,000) and multiplied by 100,000. The total New York City death rate can be used as a benchmark; in other words, it is the expected rate assuming that COVID-19 deaths affect all population groups equally. Death rates are considered "unadjusted" because they do not account for any other differences in demographic characteristics of the race/ethnicity groups, such as age (discussed in greater detail below).

This overall death rate can be disaggregated by race/ethnicity. For instance, the COVID-19 death rate for the Black/African-American population in New York City can be calculated as the number of COVID-19 deaths among the Black/African-American population (5,500) divided by the size of the Black/African-American population (1,850,000) and multiplied by 100,000. Comparing the COVID-19 death rate in the New York City Black/African-American population to the overall New York City rate—300 and 240 deaths per 100,000 population, respectively—indicates that the Black/African-American population experienced a higher-than-expected COVID-19 death rate.

Rates are useful measures for making comparisons among groups when the sizes of the groups are different, as is often the case for race/ethnicity groups. For instance, the COVID-19 death rates in the Black/African-American and Hispanic/Latino populations (300 and 230 deaths per 100,000 population, respectively) were higher than those of the White and Asian/Pacific Islander populations (180 and 130 deaths per 100,000 population, respectively), suggesting that Black/African American and Hispanic/Latino populations were more affected by COVID-19 deaths. Similarly, death rates are also useful in comparing COVID-19 deaths from different countries where the population sizes vary (e.g., the United States and Italy).

As with counts, the substantial number of deaths with unknown race/ethnicity affect the usefulness of death rates. When a substantial group is missing demographic data, the relative size of the rates may be more relevant than the actual values of the rates. Specifically, we can say with more confidence that the Black/African-American population had a higher-than-expected COVID-19 death rate, compared with saying that the Black/African-American population had 300 deaths per 100,000 population. Similar to counts, death rates for known race/ethnicity groups are likely underestimated without accounting for those with unknown race/ethnicity. For this reason, collecting complete demographic data is essential to understanding which groups are disproportionally affected, and to what degree.

3.3.4 Age-adjusted death rates

Age-adjusted death rates improve on unadjusted death rates by holding constant the age distributions between the population groups. According to the Pew Research Center, racial and ethnic minority groups in the United States tend to be younger than the White population.[41] Age-adjustments make a difference when death rates vary by age as they do for COVID-19, from which those in older age groups are more likely to die.[42] In contrast, age-adjustments make less of a difference in situations where deaths are less correlated with age. The purpose of the age-adjustment is to hold the age distributions constant so that analysts can focus on other demographics, such as race/ethnicity, without being concerned that differences are due to the different age distributions of the race and ethnic groups. Age-adjustments are particularly important for making comparisons between countries with different population distributions—such as between the United States and, for example, Italy, which has a greater proportion of the population that is age 65 and over compared to the United States.[43]

Calculating age-adjusted death rates requires knowing the age distribution of deaths within each population group of interest. For example, to calculate the age-adjusted death rate for the New York City Black/African-American population, we would need to know the age distribution of the 5,500 deaths among this group.

For New York City, compared to the unadjusted death rates, age-adjustment reduces the COVID-19 death rate in the Black/African-American and White populations as of May 13, 2020, to 270 and 130 deaths per 100,000, respectively, and increases the rate in the Hispanic/Latino population to 260 deaths per 100,000 (see fig. 6). The death rate for the Asian/Pacific-Islander population remains about the same. The ranking of the rates remains the same

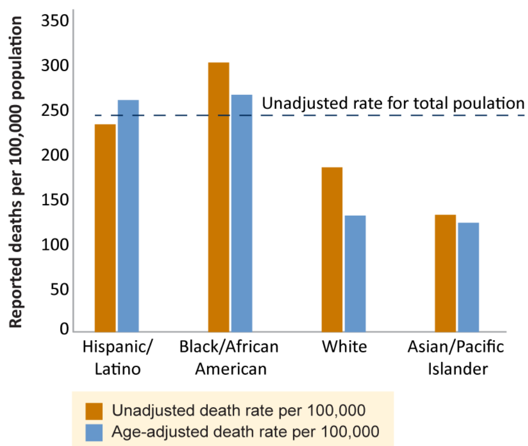
[41] Katherine Schaeffer, "The Most Common Age among Whites in U.S. Is 58—More Than Double That of Racial and Ethnic Minorities," Fact Tank, Pew Research Center, accessed June, 23, 2020, https://www.pewresearch.org/fact-tank/2019/07/30/most-common-age-among-us-racial-ethnic-groups/.

[42] For CDC data on deaths by age, see Centers for Disease Control and Prevention, Daily Updates of Totals by Week and State, accessed June 30, 2020, https://www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm.

[43] World Bank Group, *Population Ages 65 and above (% of Total Population*, accessed June 28, 2020, https://data.worldbank.org/indicator /SP.POP.65UP.TO.ZS.

after age-adjustment, with the Black/African-American and Hispanic/Latino populations having the highest death rates. Similar to unadjusted death rates, the substantial number of deaths with unknown race/ethnicity affects the usefulness of age-adjusted death rates.

Figure 6: COVID-19 reported death rates by race/ethnicity in New York City, as of May 13, 2020



Source: GAO analysis of New York City Department of Health and Mental Hygiene data. | GAO-20-635SP

Note: Hispanic/Latino includes people of any race. We cannot calculate the rate for those with Other or Unknown race/ethnicity.

3.4 Additional analytical approaches can help address imperfect COVID-19 reporting

As discussed earlier in this report, COVID-19 deaths are subject to incomplete and inconsistent reporting. Two methods can provide additional insight to help fill the gap: (1) examining deaths due to respiratory diseases and (2) examining higher-than-expected deaths from all causes during the COVID-19 pandemic. Higher-than-expected deaths are also known as excess deaths.

3.4.1 Examining deaths due to respiratory diseases

Examining deaths due to respiratory diseases—including pneumonia and influenza, as well as COVID-19—can help to account for some COVID-19 deaths that may have been misclassified. Specifically, COVID-19 deaths may have been erroneously classified as pneumonia or influenza deaths on death certificates. Even after awareness of the pandemic became more widespread, COVID-19 deaths may have been classified as pneumonia or influenza in the absence of a positive test result due to limited COVID-19 testing.

CDC's National Center for Immunization and Respiratory Diseases publishes the FluView-Interactive surveillance system, which includes a comparison of the percentage of all deaths that are due to influenza or pneumonia to a seasonal baseline based on NCHS death certificate data.[44] This surveillance system releases data on pneumonia and influenza deaths 1 week after the week of death, in order to collect enough data to produce statistically reliable estimates of the percentage of all deaths due to each cause. Adjusting for seasonality, the surveillance system detected higher-than-expected pneumonia and influenza deaths in the week ending February 29, 2020, while according to provisional death counts, only five COVID-19 deaths were detected for that week. Some of these higher-than-expected pneumonia and influenza deaths may have

[44]For more information, see Centers for Disease Control and Prevention, *U.S. Influenza Surveillance System: Purpose and Methods*, accessed July 21, 2020, https://www.cdc.gov/flu/weekly/overview.htm

been COVID-19 cases that were misclassified. Examining deaths due to respiratory diseases can help analysts to better understand the full scope of COVID-19 deaths, particularly in the early stages of the pandemic before guidance was available on classifying COVID-19 deaths.

3.4.2 Examining higher-than-expected deaths during the pandemic

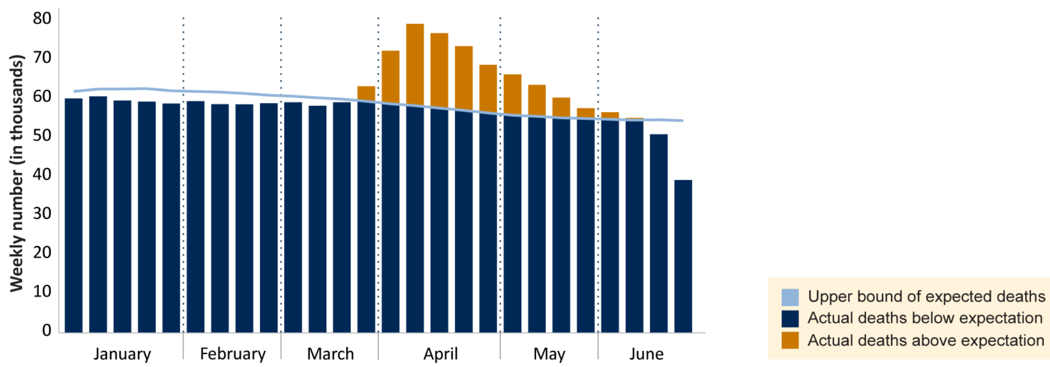
Examining higher-than-expected deaths from all causes that have occurred during the COVID-19 pandemic also helps to address the imperfect reporting of COVID-19 deaths, because the number of total deaths is likely more accurate than the numbers of deaths from specific causes. This analytic method is also known as examining excess deaths. Excess deaths are calculated by comparing the number of deaths from all causes to the number of deaths that would typically be expected during a specific time period. The number of deaths from all causes is likely more accurate than the number of deaths classified by specific cause because the cause of death may be unknown or misdiagnosed. Therefore, using the number of deaths from all causes in an analysis of excess deaths may mitigate some of the data limitations discussed earlier.

An NCHS report on excess deaths associated with COVID-19 found that significant excess deaths occurred nationally after the week ending in March 28, 2020. Figure 7 shows total deaths from all causes from the week ending January 4, 2020, through June 27, 2020 (data as of July 14, 2020), compared with the upper range of the expected variation in mortality. Deaths above this threshold are considered excess deaths. Excess deaths have been used in other settings to estimate deaths related to a natural disaster. For example, at least three groups of academic researchers estimated excess deaths in Puerto Rico associated with Hurricane Maria, where there were challenges identifying and documenting deaths related to the hurricane.[45]

Excess deaths help account for the imperfect reporting of COVID-19 deaths, but there are some important considerations.

[45]See GAO, *Disaster Response: Federal Assistance and Selected States and Territory Efforts to Identify Deaths from 2017 Hurricanes*, GAO-19-486 (Washington, D.C.: Sept. 13, 2019).

Figure 7: Data on higher-than-expected weekly mortality for 2020, as of July 14, 2020



Source: GAO analysis of Centers for Disease Control and Prevention (CDC)/National Center for Health Statistics (NCHS) data.. | GAO-20-635SP

Notes: The weekly total provisional deaths from all causes are as of July 14, 2020. Provisional deaths are incomplete, especially for the most recent weeks. NCHS uses historical data and complex statistical processes to estimate the expected number of deaths in a given state and week, accounting for seasonal patterns. The number of excess deaths is calculated as the difference between the observed number of deaths and the expected number, based on the upper bound of the 95 percent confidence interval of the average expected number of deaths. Observed numbers of deaths are weighted to partially account for incomplete reporting of deaths in the most recent weeks. Further details on the methods may be found at https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm.

- Excess deaths are not necessarily deaths caused by COVID-19 and may include deaths caused both by COVID-19 and other causes. For instance, deaths indirectly related to COVID-19 may include deaths due to other causes because of health care shortages.

- Excess deaths are sensitive to how the expected number of deaths is estimated. For example, one method is to average deaths from all causes during the same week for prior years (e.g., the period 2017–2019), which also accounts for the seasonality of deaths. Another method is modeling, which can account for additional factors, such as a population's age and sex. Different definitions of expected deaths will have an effect on the magnitude of excess deaths.

- Delays in reporting deaths from all causes can range from 1 to 8 or more weeks after deaths occur. Data in recent weeks are more subject to being incomplete. NCHS takes steps to account for this delay in its estimates of excess deaths, but the adjusted estimates are still likely to be incomplete.

3.5 Comparing deaths from COVID-19 and other causes can provide context when data become available

Comparing the ranking of the number of COVID-19 deaths with the number of deaths from other leading causes provides context to help understand the magnitude and relative burden of COVID-19 deaths. These rankings are also used to guide the setting of public health priorities. However, the rankings are limited in that 2020 death data are provisional and will not be finalized until the end of 2021.

Final data on deaths will be available 11–12 months after the end of the data year. For instance, 2020 final data are expected to be published in November or December 2021. Provisional mortality data are typically available 3–9 months after deaths occur. However, during the current pandemic, provisional counts for COVID-19 deaths, total deaths from all causes, and deaths from several other causes are being released 1–2 weeks after deaths occur.

Without final data on leading causes of death, one approach is to use leading causes of death from prior years. Some analyses average the death counts from leading causes for a given time period from the recent years for each leading cause of death to obtain a more reliable estimate of expected deaths from leading causes. However, estimates of leading causes of death from prior years may not be an accurate stand-in for the leading causes of death in 2020, if COVID-19 has affected those leading causes. NCHS has started releasing provisional data on some causes of death, which may provide an initial indication of how COVID-19 deaths will rank among the leading causes. However, for more accurate results, analysts will need to revisit this analysis when provisional counts of leading causes of death and final counts of death in 2020 are available.

4 Forecasting Models Can Provide Valuable Insights, but Understanding Their Purpose and Limitations Is Essential for Interpreting Results

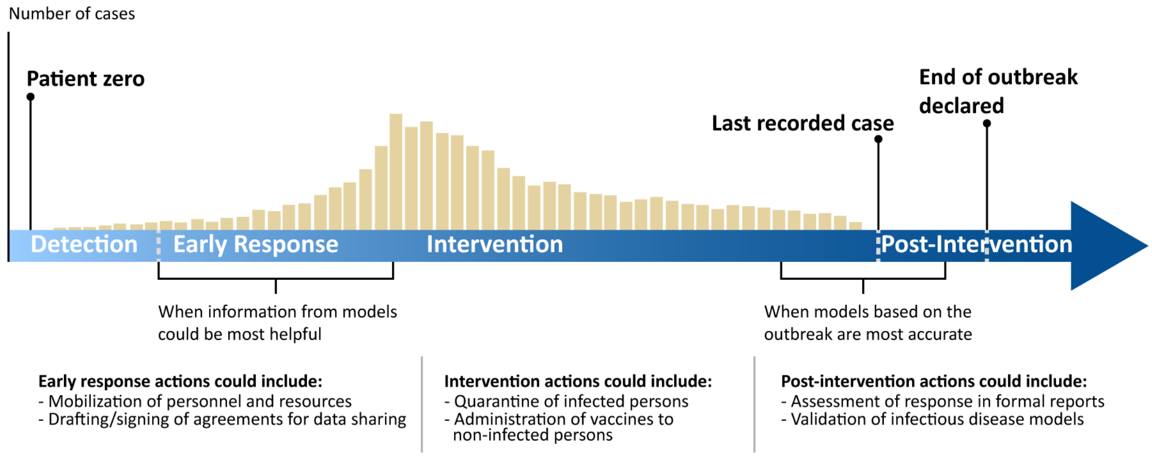
4.1 Model results can vary for many reasons, and predictions are likely to be less accurate and less precise early in an outbreak

Models are one tool used by public health experts to study and predict the spread of infectious diseases. In response to the COVID-19 pandemic, various models have been used to estimate the disease's effect on total deaths, deaths per day, the timing of peak cases and deaths, and hospital resources.

Model results can vary based on the model's uses, purpose, data inputs, and assumptions, among other factors. For example, if the purpose of a model is to present scenarios that illustrate the effectiveness of different ways to mitigate the epidemic, the scenarios will be driven by those options and results may be unrealistic if the scenarios are not followed. Although the model's purpose was never to project what is most likely to occur, its results may be overly simplified, and communications to the public by various entities may misrepresent some scenarios as projections of what will occur. Model developers may also choose different data inputs, such as different values for how infectious a disease is (estimates for the current pandemic vary). In addition, a model may be based on prior information from similar diseases or past outbreaks, current data on the disease, or both. Even if two models have the same purpose and receive the same data inputs, they may still produce different results if their underlying assumptions differ. For example, one model may assume everyone has an equal chance of getting the disease, while another may assume transmission is more or less likely based on age, where people live and work, or factors related to the disease itself. To understand a model's results, such assumptions must also be understood.

Early in an outbreak, knowledge of the disease may be limited, and many assumptions used in early models are generally based on the dynamics of the infection where it first appeared. These assumptions may not hold up over time, as the dynamics of the infection may vary between locations and populations or change as the disease progresses. However, as data and knowledge of the disease improve, the model's predictions can become more accurate and precise. Figure 8 illustrates how during the outbreak of a new disease, models can be most helpful early in the response, but are most limited by a lack of data. Later in the outbreak, more data become available, but there is less time to implement an optimal response for ending the outbreak.

Figure 8: **Timeline of epidemic outbreak versus timeline of model utility and accuracy**



Source: GAO analysis of documents. | GAO-20-635SP

Infectious disease modeling in general faces the following challenges that may make results less precise, less accurate, and harder to explain:

- **Data.** Models are inherently limited by the underlying data, which can be scarce and inconsistent, especially early in a novel disease outbreak, lessening the precision and accuracy of model estimates. As we discussed in chapter 2, there are several limitations in CDC-reported data—including cases, hospitalizations, and deaths—that are commonly used in COVID-19 modeling.

- **Uncertainty.** Especially in the early stages of a novel infectious disease outbreak, there are multiple sources of uncertainty, such as overreliance on a few data points and lack of information on behavioral changes that might occur among the population. For example, we discussed earlier that CDC-reported data on COVID-19 cases and deaths may be incomplete and may not capture the full effect of COVID-19 due to limited testing availability and the potentially high number of infected people who may be asymptomatic, among other issues.

- **Communication.** Models can be complex and thus difficult to explain to lay audiences. As a result, the methods, outputs, and even the purpose of a model may be oversimplified and thus misunderstood. For example, the projections of a model will often initially be published with a range of uncertainty, but communications to the public by various entities may not report results with the full range of uncertainty and may report only the highest number.

4.2 Mechanistic and statistical models have different purposes and limitations

There are two broad categories of infectious disease models—mechanistic and statistical:

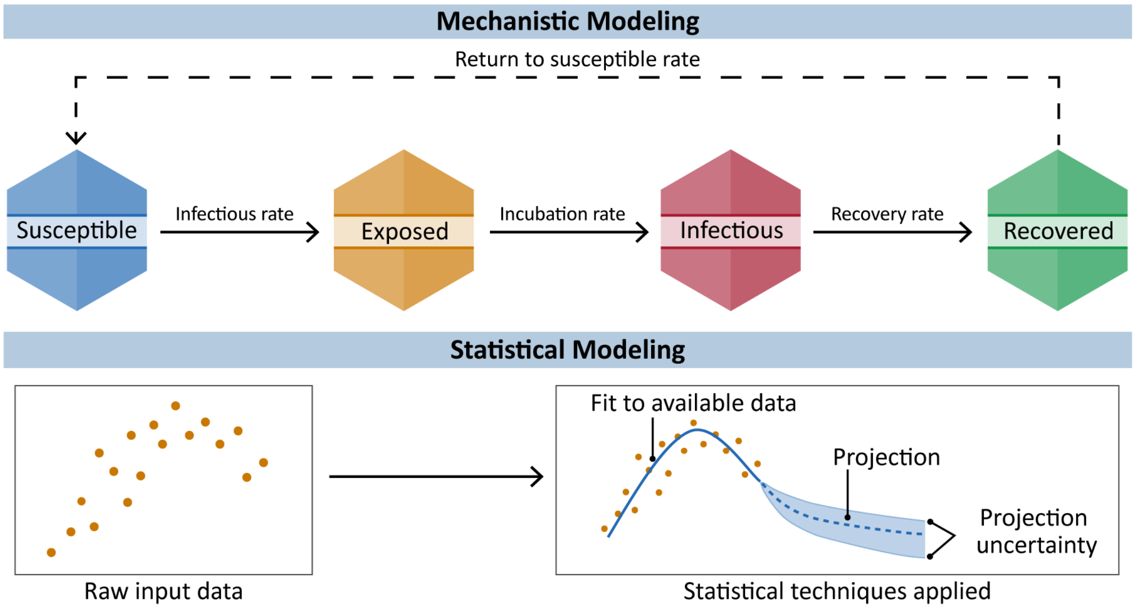
- **Mechanistic models** use equations, based on scientific understanding of disease dynamics and human behavior, to represent the mechanics of how a disease progresses. An example of a commonly used mechanistic model in infectious disease modeling is the Susceptible-Exposed-Infectious-Recovered (SEIR) model, which describes how a population moves from being vulnerable to infected through the stages of a disease. For example, movement from susceptible to exposed can be based on the infectiousness of the pathogen and the degree of contact between infected individuals and susceptible individuals.[46] With mechanistic models, researchers can estimate the effect of proposed interventions, such as social (i.e., physical) distancing, by adjusting the model and rerunning it.

- **Statistical models,** by contrast, use only data on a current or past outbreak (such as the number of cases, hospitalizations, deaths, or a combination of these) and not scientific understanding of the disease. For example, they might use data on reported deaths to develop an equation to forecast future deaths or hospital needs. Over the past century, the statistical approach has become increasingly popular among epidemiologists. More recently, the increase in computing power available to researchers has encouraged modelers who prefer the mechanistic approach to add statistical components to their models. Figure 9 illustrates how mechanistic and statistical models work.

Based on their design, assumptions, and strengths, mechanistic models and statistical models can be suited to different purposes, which can affect the results. For example, because a mechanistic model relies on knowledge of the disease rather than only on current data, it can be used to develop hypothetical scenarios that attempt to show the effect of a decision on the course of an epidemic. However, such a model may only compare intervention scenarios against an unlikely worst-case scenario in which nothing is done to respond to an outbreak and the population does not change behavior, even at the height of the epidemic. In such a case, it is important for decision makers to understand all of the assumptions underlying a “do nothing” scenario to evaluate the extent to which it can be a useful tool for comparison.

[46]A subset of the mechanistic approach known as agent-based or individual simulation models can also be used to model disease spread. For example, some models use simulations of interactions between infected and noninfected populations to better inform their results. These simulations use preprogrammed rules derived from scientific understanding of disease dynamics and human behavior to simulate the movement, interaction, and disease status of individuals in a single hospital, a city, or even the entire United States.

Figure 9: **Mechanistic versus statistical modeling**



Source: GAO illustration of standard mechanistic modeling (top); GAO illustration of standard statistical modeling (bottom). GAO-20-635SP

Similarly, a statistical model that is intended to forecast cases based only on current and prior data may not include expected developments that could affect the course of the outbreak, such as the relaxation of social distancing in different locations. Such a model could be updated when more data are available on the effects of those developments, and these updates could affect model results. Users of the statistical model would need to understand that it should be updated frequently to account for real-world developments that could affect the data the model uses.

Table 3 describes two examples of models that have had some prominence in briefings by policy makers and the media, and for which the model description and other details were readily available—the Institute for Health Metrics and Evaluation (IHME) model and the Imperial College London model. We reviewed information on their assumptions, methodology, and main findings. We did not compare projections between the IHME and Imperial College London models because the models make different assumptions, report over different time frames, and have different purposes. The information presented for the Imperial College London model is based on the initial March 16, 2020 version.[47] For the IHME model, the table presents the model version as of May 4, 2020, when IHME made major updates to its model.[48]

[47]We chose this model for comparison from among other mechanistic models based on media reports that it may have had a major influence on government policy toward COVID 19, particularly in the United Kingdom.

[48]Some details are documented in an IHME April 26, 2020, preprint (not peer-reviewed) paper; see IHME COVID-19 Health Service Utilization Forecasting Team, "Forecasting the Impact of the First Wave of the COVID-19 Pandemic on Hospital Demand and Deaths for the USA and European Economic Area Countries," preprint, submitted April 26, 2020, https://www.medrxiv.org/content/10.1101/2020.04.21.20074732v1.

Table 3: Key features and description of the Imperial College London and Institute for Health Metrics and Evaluation (IHME) models

	Imperial College London model (as of March 16, 2020)	IHME model (as of May 4, 2020)
What is it	This is a mechanistic model based on two earlier pandemic influenza models. It simulates the COVID-19 outbreak in the United States and United Kingdom, and estimates the effects of interventions in both countries.	This is a hybrid model that blends different approaches (statistical and mechanistic) to improve forecasts as new data and information on the disease become available.
How does it work?	This model simulates the spread of the disease in the United States and Great Britain using population data and simulated transmission in households, schools, workplaces, and the wider community.	The IHME model has a statistical component for death rates, a Susceptible-Exposed-Infectious-Recovered (SEIR) component for disease transmission, and a simulation of hospital use. The statistical component fits an equation to published deaths from around the world (including the United States), while the SEIR component incorporates information on disease dynamics, such as the basic COVID-19 factors (like the R_t) and other factors that can affect transmission (such as social distancing and temperature).[a] The hospital simulation estimates future usage of hospital resources based on local usage patterns (where available) or pooled data from other U.S. states (where local data are not available).
What does it predict?	This model predicts the effect of two strategies of intervention, including a "do nothing" scenario and combinations of interventions, on cases, deaths, and health care demand, according to various mitigation and suppression goals at the national level. It projects total deaths by October 2020.	This model predicts daily deaths; infections; testing; need for beds, ventilators, and intensive care units; and total deaths though August 4.
What data inputs does it use?	This model uses the early growth-rate data of the epidemic in Wuhan, China, to establish how infectious the disease is, and census, school, and workplace data to simulate how the virus spreads.	The statistical component uses local government, national government, academic and nonprofit, and World Health Organization websites as sources of death rate data at the state or province level. The SEIR component models disease transmission as a function of mobility, air temperature, testing rates, and the proportion of populations who live in dense areas. The hospital simulation inputs were based on a prior study in the United States and data from the United States and European Economic Area countries obtained during the current outbreak.
What key assumptions does it use?	The model applies interventions (e.g., social distancing and quarantine) across	For locations without formal quarantine policies, the model assumes social

	Imperial College London model (as of March 16, 2020)	IHME model (as of May 4, 2020)
	the United States, without variation in the timing or implementation by states. In the unmitigated ("do nothing") scenario, asymptomatic individuals do not change behavior over the course of the epidemic.[b]	distancing is occurring if a 40 percent decline in population mobility is observed (based on data from private sector companies). Where states have eased social distancing, the model assumes it is no longer occurring. Where states have not eased social distancing, the model assumes it will run until August 4, 2020. The SEIR component relies on reported daily deaths by location and estimated infection-fatality ratios to infer the size of the infected population in locations and estimate future cases.[c] Ambient temperature, testing rates, and population density are additional factors driving transmission.
How and why do the assumptions matter?	If individuals do change behavior voluntarily, they may reduce transmission without interventions, which makes the effectiveness of the strict policy interventions appear larger than it may actually be.	The core of the IHME model is the statistical component, which relies on reported data on COVID-19 deaths. Therefore, substantial variance in the quality and availability of the data will affect the accuracy of the model. It is also possible that in states where social distancing has eased, segments of the population may keep practicing social distancing, despite the lack of a policy.

Source: GAO analysis of literature sources including Imperial College COVID-19 Response Team, "Report 9: Impact of Non-pharmaceutical interventions (NPIs) to reduce COVID-19 Mortality and Healthcare Demand," available online March 16, 2020; IHME COVID-19 Health Service Utilization Forecasting Team, " "Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by U.S. state in the next 4 months," preprint, published March 26, 2020; IHME COVID-19 team,"COVID-19: What's New for May 4, 2020," http://www.healthdata.org/sites/default/files/files/Projects/COVID/Estimation_update_050420.pdf; Rahi Abouk and Babak Heydari, "The Immediate Effect of COVID-19 Policies on Social Distancing Behavior in the United States," https://www.medrxiv.org/content/10.1101/2020.04.07.20057356v2, preprint, posted April 28, 2020. | GAO-20-635SP

[a]R_t refers to the effective reproduction number, which is a measure of transmissibility of the infection at a specific time 't' over the course of an epidemic.

[b]Recent research has suggested that the assumption that the U.S. populace would not voluntarily change behavior during the pandemic was unrealistic. In a study focused on the early phase of the pandemic, changes in behavior appeared to be driven by a combination of policy forces and voluntary actions, with voluntary actions and awareness-raising mechanisms at least as effective as some interventions (such as bans on large gatherings).

[c]The infection-fatality ratio is the number of people infected by a disease who die, divided by the total number of people infected by the disease.

Both of these types of models have potential advantages for policy- and decision makers:

- Mechanistic models can project outcomes such as total deaths for various scenarios of policy actions. For example, the Imperial College model assesses the effect of five different nonpharmaceutical interventions (e.g., social distancing measures), implemented individually and in combination, on the ability to achieve different levels of mitigation or suppression policy goals.[49] This approach lets a decision maker consider the relative costs and benefits of different options for addressing the pandemic. If the decision maker has the resources and the time to deploy them, modeling for multiple scenarios can be useful, because it can support preparation for a range of scenarios. If the scientific understanding of the disease used to inform their assumptions is accurate, mechanistic models may be able to make more accurate short-term projections early in an outbreak, especially if the disease is not a novel one.

- Because statistical models can base their projections on data specific to the current outbreak, their projections may be more reliable for a novel outbreak (for which there would be no preexisting scientific knowledge). Because the IHME model incorporated real-time COVID-19 data (as opposed to data from previous, similar diseases or only early information on the current epidemic), its projections may become more accurate over time as more data emerge. The statistical component of the IHME model allows for updating the model projections as new data become available. The SEIR component incorporates factors such as changing behavior in the population and new or improved information and understanding about the biology of the disease.

However, both model approaches also have limitations for use as tools in guiding policy responses to the pandemic:

- Mechanistic models may rely on accurate knowledge of disease transmission to estimate rates of disease spread, and in SEIR models, to estimate the number of people expected to ultimately be exposed, infectious, and recovered. Initially, it was thought that the COVID-19 virus would be mostly transmitted via symptoms (e.g., coughing), like the similar virus that caused the 2003 SARS outbreak (SARS-CoV-1). However, asymptomatic transmission is now known to be a major contributor to disease spread.[50] Initial estimates of the rate of transmission were thus inaccurate, because they focused on transmission by symptomatic individuals. Such limitations in the data and information about asymptomatic infection rates and their effect on spread of the virus also affect the levels of uncertainty in mechanistic modeling.

- Because statistical models are more reliant on data, factors that limit data quality and availability can limit the accuracy and precision of the model. Statistical models also have to use the recent past to project the future, which may deviate significantly as a function of population behavior or government policies in response to the ongoing pandemic. In addition, for each additional data source or factor added to a statistical model to improve predictions,

[49]Mitigation focuses on slowing but not necessarily stopping epidemic spread—reducing peak health care demand while protecting those most at risk of severe disease from infection—while suppression aims to reverse epidemic growth, reducing case numbers to low levels and maintaining that situation indefinitely.

[50]Daniel P. Oran, and Eric J. Topol, “Prevalence of Asymptomatic SARS-CoV-2 Infection: A Narrative Review,” *Annals of Internal Medicine* (June 3, 2020).

there will be additional uncertainty associated with those data, and analysts must evaluate and incorporate this uncertainty into the final model results.

Another option to improve model accuracy is to use ensemble modeling, which combines entirely disparate model forecasts together. In ensemble modeling, the results of independent models can be combined to produce an average that is often more accurate than any of the individual models. A 2018 study showed that statistical models used in the study were better at predicting longer-term dynamics of seasonal influenza (e.g., seasonal onset), while mechanistic models used in the study were better at predicting short-term dynamics (e.g., forecasts 2 weeks out), and the ensemble model outperformed both sets of models in terms of overall accuracy.[51]

Using an ensemble model, or simply having multiple models for the same disease—including an emerging infectious disease such as COVID-19—can increase confidence in projections if the models broadly agree. If the results disagree, having multiple models can help identify which modeling approaches may be better suited for a given disease and purpose, if actual data are available to compare the model projections against. For either scenario, full transparency of methods and data is essential for understanding the models and interpreting any differences among them.

[51]Sasikiran Kandula et al., “Evaluation of Mechanistic and Statistical Methods in Forecasting Influenza-like Illness,” *Journal of the Royal Society Interface*, (July 25, 2018).

5 Strategic Implications

The data that CDC reports on COVID-19 are essential in helping researchers and decision makers understand how the virus is affecting the U.S. population, and these data can inform decisions about how to respond. However, because of important limitations in consistency and completeness, the available data may not tell the full story of the effects of COVID-19—for example, cases and deaths may be underreported. Thus, it is essential to consider these limitations in using the data in the decision-making process.

However, various analytical approaches can help researchers and decision makers use the available data to understand the dynamics of COVID-19 and its effects. In particular, some methods can help those studying COVID-19 to understand how it is affecting groups of people differently, and other methods can provide insights into the magnitude of COVID-19 and can address some of the issues that affect the quality of available data. Regardless, it is important for researchers and decision makers to understand both the uses and limitations of these analytical approaches in order to use them effectively.

Finally, forecasting models have been widely used to help predict how COVID-19 will evolve, and models can be an important part of an analytical strategy. But, models are inherently limited by the quality and completeness of the data used to build them. Thus, they are less likely to be accurate early in an epidemic, when data quality and availability can be limited. Models are also complex and easily misunderstood, and therefore decision makers must understand their features, purposes, and limitations in order to use them appropriately.

6 Agency and Expert Comments

We provided a draft of this report to the Secretary of the Department of Health and Human Services (Centers for Disease Control and Prevention) and to an external expert, with a request for technical comments. We incorporated comments into this report as appropriate.

We are sending copies of this report to the appropriate congressional committees, relevant federal agencies, and other interested parties. In addition, the report is available at no charge on the GAO website at http://www.gao.gov.

If you or your staff members have any questions about this report, please contact Timothy M. Persons at 202-512-6888 or PersonsT@gao.gov, SaraAnn Moessbauer at 202-512-4943 or MoessbauerS@gao.gov, or Mary Denigan-Macauley at 202-512-7114 or DeniganMacauleyM@gao.gov.

Contact points for our Offices of Congressional Relations and Public Affairs may be found on the last page of this report. GAO staff who made key contributions to this report are listed in appendix II.

T.M. Persons

Timothy M. Persons, PhD, Chief Scientist and Managing Director,
Science, Technology Assessment, and Analytics

SaraAnn W. Moessbauer

SaraAnn Moessbauer, Director,
Applied Research and Methods

Mary Denigan-Ma

Mary Denigan-Macauley, PhD, Director,
Health Care

Congressional Addressees

The Honorable Richard C. Shelby
Chairman
The Honorable Patrick J. Leahy
Vice Chairman
Committee on Appropriations
United States Senate

The Honorable Lamar Alexander
Chairman
The Honorable Patty Murray
Ranking Member
Committee on Health, Education, Labor, and Pensions
United States Senate

The Honorable Ron Johnson
Chairman
The Honorable Gary C. Peters
Ranking Member
Committee on Homeland Security and Governmental Affairs
United States Senate

The Honorable Nita M. Lowey
Chairwoman
The Honorable Kay Granger
Ranking Member
Committee on Appropriations
House of Representatives

The Honorable Frank Pallone, Jr.
Chairman
The Honorable Greg Walden
Republican Leader
Committee on Energy and Commerce
House of Representatives

The Honorable Bennie Thompson
Chairman
The Honorable Mike Rogers
Ranking Member
Committee on Homeland Security
House of Representatives

The Honorable Carolyn B. Maloney
Chairwoman
The Honorable James Comer
Ranking Member
Committee on Oversight and Reform
House of Representatives

Appendix I: Objectives, Scope, and Methodology

This report discusses (1) collection methods and limitations of the Coronavirus Disease 2019 (COVID-19) surveillance data for cases, hospitalizations, and deaths reported by the Centers for Disease Control and Prevention (CDC),[1] (2) approaches for analyzing COVID-19 data, (3) uses and limitations of modeling for understanding COVID-19.

To identify available data sources on COVID-19 and their limitations, we reviewed data reported by CDC and state health departments. We obtained documentation on CDC's surveillance systems and efforts to report data on cases, deaths, and hospitalizations from its website, reports, and other publications. For cases, we reviewed documentation from CDC and the Council of State and Territorial Epidemiologists websites and associated reports and publications regarding COVID-19 cases and the standardized case definition. For hospitalizations, we reviewed data and documentation on CDC's COVID-19 hospitalizations surveillance system (COVID-NET) website, including published reports. For deaths, we obtained and reviewed documentation on surveillance of COVID-19 deaths using death certificate data from CDC's National Center for Health Statistics (NCHS) website and documentation and data from death reporting through the case surveillance system. Specialist staff with training in public health surveillance, epidemiology, and biostatistics reviewed the documentation to identify the processes used to define variables, collect and validate data, and develop public reports. After assembling documentation to describe the data collection processes, specialist staff identified limitations of the available data and considerations for analysis by applying expertise in public health surveillance data collection and analysis.

To identify methods of analyzing data on COVID-19 and their uses and limitations, we reviewed available data and applied staff expertise in public health, epidemiology, and biostatistics. We obtained data on COVID-19 cases from CDC; excess deaths related to COVID-19 from CDC's NCHS; COVID-19 cases, hospitalizations, and deaths from the Minnesota public health department; and demographics of COVID-19 decedents from the New York City Department of Health and Mental Hygiene. We assessed the reliability of these data by reviewing documentation on their sources and collection methods and the consistency of trends over time for the purpose of comparing methods of analysis. We used data on excess deaths from an NCHS report after comparing the report's methods against accepted practices for excess death analysis in epidemiology.[2]

To describe uses and limitations of COVID-19 forecasting models, we leveraged prior GAO reports that identified the goals and methods of disease modeling, in general, and focused on two prominent forecasting models for COVID-19, in particular. We also used prior

[1]This report does not include other types of surveillance efforts conducted by CDC and some state and local health officials, such as COVID-19 outpatient and emergency department illness surveillance or recovered cases surveillance; however, we are examining these other types of data for our future work.

[2]The report and data on excess deaths are available at https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm (accessed June 15, 2020).

GAO work on disease modeling and specialist staff expertise to identify the purpose, structure, input data, and results of forecasting models.[3] Prior GAO reports used expert and agency interviews, searches of available published literature, and review of agency documents to develop background material on modeling in general and infectious disease modeling in particular. To describe how these general principles of disease modeling apply to COVID-19, we gathered in-depth documentation on the models developed by Imperial College London and the Institute for Health Metrics and Evaluation. We selected these models based on their use of mechanistic and statistical methods, their frequent mentions in news media reports, and their reported use by policymakers, such as the use of Imperial College London model by the government of the United Kingdom to inform decisions on social distancing policy and COVID-19.

We conducted our work from May 2020 to July 2020 in accordance with all sections of GAO's Quality Assurance Framework that are relevant to technology assessments. The framework requires that we plan and perform the engagement to obtain sufficient, appropriate evidence to meet our stated objectives and to discuss any limitations to our work. We believe that the information and data obtained, and the analysis conducted, provide a reasonable basis for any findings and conclusions in this product.

[3] GAO, *Infectious Disease Modeling: Opportunities to Improve Coordination and Ensure Reproducibility*, GAO-20-372, (Washington, D.C.: May 13, 2020).

Appendix II: GAO Contacts and Staff Acknowledgments

GAO Contacts

Timothy M. Persons, PhD at 202-512-6888 or PersonsT@gao.gov

SaraAnn Moessbauer at 202-512-4943 or MoessbauerS@gao.gov

Mary Denigan-Macauley, PhD at 202-512-7114 or DeniganMacauleyM@gao.gov.

Staff Acknowledgments

In addition to the contacts named above, the following individuals made contributions to this report: Karen Howard (Director), Jeff Tessin (Assistant Director), Edward J. Rice (Analyst-in-Charge), Mariel Alper, Jennifer Beddor, Jehan Chase, Karen Doran, Lorraine Ettaro, Justin Fisher, Tom Lombardi, Anika McMillon, Jennifer Schwartz, Amber Sinclair, Frances Tirado, Jack Wang, and Sirin Yaemsiri.

(104333)

Related GAO Products

COVID-19: Opportunities to Improve Federal Response and Recovery Efforts. GAO-20-659T. Washington, D.C.: June 26, 2020.

COVID-19: Opportunities to Improve Federal Response and Recovery Efforts. GAO-20-625. Washington, D.C.: June 25, 2020.

Science & Tech Spotlight: COVID-19 Modeling. GAO-20-582SP. Washington, D.C.: June 4, 2020.

Science & Tech Spotlight: COVID-19 Vaccine Development. GAO-20-583SP. Washington, D.C.: May 26, 2020.

Science & Tech Spotlight: COVID-19 Testing. GAO-20-584SP. Washington, D.C.: May 20, 2020.

Infectious Disease Modeling: Opportunities to Improve Coordination and Ensure Reproducibility. GAO-20-372. Washington, D.C.: May 13, 2020.

Science & Tech Spotlight: Social Distancing During Pandemics. GAO-20-545SP. Washington, D.C.: May 13, 2020.

Science & Tech Spotlight: Coronaviruses. GAO-20-472SP. Washington, D.C.: March 3, 2020.

Emerging Infectious Diseases: Actions Needed to Address the Challenges of Responding to Zika Virus Disease Outbreaks. GAO-17-445. Washington, D.C.: May 23, 2017.

GAO's Mission

The Government Accountability Office, the audit, evaluation, and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through GAO's website (https://www.gao.gov). Each weekday afternoon, GAO posts on its website newly released reports, testimony, and correspondence. To have GAO e-mail you a list of newly posted products, go to https://www.gao.gov and select "E-mail Updates."

Order by Phone

The price of each GAO publication reflects GAO's actual cost of production and distribution and depends on the number of pages in the publication and whether the publication is printed in color or black and white. Pricing and ordering information is posted on GAO's website, https://www.gao.gov/ordering.htm.

Place orders by calling (202) 512-6000, toll free (866) 801-7077, or
TDD (202) 512-2537.

Orders may be paid for using American Express, Discover Card, MasterCard, Visa, check, or money order. Call for additional information.

Connect with GAO

Connect with GAO on Facebook, Flickr, Twitter, and YouTube.
Subscribe to our RSS Feeds or E-mail Updates.
Listen to our Podcasts and read The Watchblog.
Visit GAO on the web at https://www.gao.gov.

To Report Fraud, Waste, and Abuse in Federal Programs

Contact: Website: https://www.gao.gov/fraudnet/fraudnet.htm
Automated answering system: (800) 424-5454 or (202) 512-7470

Congressional Relations

Orice Williams Brown, Managing Director, WilliamsO@gao.gov, (202) 512-4400, U.S. Government Accountability Office, 441 G Street NW, Room 7125, Washington, DC 20548

Public Affairs

Chuck Young, Managing Director, youngc1@gao.gov, (202) 512-4800
U.S. Government Accountability Office, 441 G Street NW, Room 7149
Washington, DC 20548

Strategic Planning and External Liaison

James-Christian Blockwood, Managing Director, spel@gao.gov, (202) 512-4707
U.S. Government Accountability Office, 441 G Street NW, Room 7814, Washington, DC 20548