ASPE ASSISTANT SECRETARY FOR PLANNING AND EVALUATION
OFFICE OF SCIENCE & DATA POLICY

ISSUE PAPER

March 2022

Quality of Federal COVID-19 Data

KEY POINTS

- HHS has publicly released a large volume of data on COVID-19 testing, cases, hospitalizations, deaths, and vaccinations, including from a new reporting system that captures daily information from the nation's hospitals.
- Applying the *Framework for Data Quality* developed by the Federal Committee on Statistical Methodology, mortality data from the National Center for Health Statistics (NCHS) are the highest quality among all HHS released COVID-19-related indicators due to their completeness and granularity.
- NCHS' excess mortality estimates measure the full extent of the pandemic's mortality impacts.
- Testing data are the lowest quality COVID-19-related data, due to the absence of data from certain care settings and the lack of granular data on patient characteristics.
- The publication of differing estimates for COVID-19 indicators by CDC, NCHS, and HHS is a potential source of confusion. To build credibility with the public, HHS should minimize and clearly explain any differences in published COVID-19 indicators.
- Inconsistencies between daily COVID-19 case and death estimates published by CDC and several non-government sources highlight the challenges of acquiring timely and accurate data from states, territories, and localities and communicating that information to the public.

I. BACKGROUND

Effective January 27, 2020, the U.S. declared a public health emergency (PHE) to support the nation's response to COVID-19.[1] The ease with which the virus spreads, coupled with the ability of asymptomatic people to transmit the virus, led to a significant worldwide pandemic in 2020 that continued into 2021 and 2022. As of March 2022, more than 960,000 Americans had died from COVID-19 and there have been more than 79 million confirmed cases.[2]

Reliable, high-quality information is critical for sound decision-making, and the pandemic highlights the importance of accurate and timely national data to inform actions, such as decisions on staffing and supplies for hospitals and healthcare facilities, the development of guidelines for workplace safety, and strategies for preventing, detecting, and responding to the current and future public health threats. The provision of high quality data is key to ensuring that decisions are made using the most accurate information possible and that decision-makers at different levels of government (local, state, territorial, and federal) can compare data across sources and time periods when responding to the pandemic.

[1] https://www.phe.gov/emergency/news/healthactions/phe/Pages/2019-nCoV.aspx
[2] https://covid.cdc.gov/covid-data-tracker/#datatracker-home

Data related to the incidence, prevalence, and outcomes of COVID-19 are used to understand the disease and to inform critical policy and program decisions on pandemic response and recovery. The importance of assessing the quality of this data cannot be overstated. Given the tangible consequences of these decisions, understanding the quality and appropriate use of the data are essential.

Data on COVID-19 in the U.S. population come from multiple sources and systems, some of which have changed over time, along with changes in case definitions and reporting requirements. Any potential issues with data quality – such as missing documentation, lack of transparency, or delayed data – mean that policy makers might be making significant decisions without a complete picture of the situation. Despite the recognition that high-quality data are needed, data quality is not always assessed. This issue brief examines the following questions:

- What data sources does HHS rely upon for information on COVID-19?
- Where and how does HHS share this information with the public?
- What data quality issues are associated with each source of data?
- What caveats must be considered in using these data for decision-making or reporting?
- How can data quality be improved and data quality issues be communicated?

II. APPROACH

This analysis first identified requirements and standards for data quality, which provide policy and procedural guidance for agencies to adopt a basic standard of quality for information they disseminate.[3] We also evaluated tools for assessing data quality, including the Bureau of Labor Statistics quality assessment tool[4] and the 2020 Federal Committee on Statistical Methodology's (FCSM) *Framework for Data Quality,*[5] which we chose as the primary framework for this assessment.

The FCSM *Framework for Data Quality* definition of data quality is informed by the Information Quality Act (IQA).[6] This framework defines data quality as "the usefulness and credibility of data and products derived from data (e.g., statistics, analyses, and visualizations). Data and data products have high quality when they capture desired information using scientifically appropriate methods to represent reality in a manner that sustains public trust." Office of Management and Budget (OMB) guidelines on implementing the IQA explain that quality encompasses utility, objectivity, and integrity, which are reflected as domains in this framework (Figure 1). In addition, updated guidance on the IQA emphasizes "fitness-for-purpose" – assessing the quality of information based on the likely use of that information – for information destined for a higher-impact purpose, noting that this information must be held to higher standards of quality.[7] Information on COVID-19

---

[3] OMB. 2002. Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies. (67 FR 8452, Feb 22, 2002). Available at: https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/assets/OMB/fedreg/reproducible2.pdf

[4] Bureau of Labor Statistics. *Development of a Quality Framework and Quality Indicators at the Bureau of Labor Statistics*. Available at: https://www.bls.gov/osmr/research-papers/2014/pdf/st140050.pdf

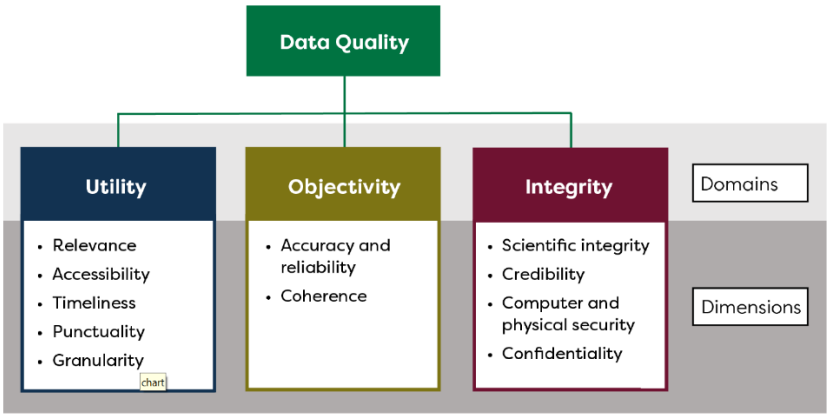
[5] FCSM-20-04. *A Framework for Data Quality.* Federal Committee on Statistical Methodology. Sept. 2020. Available at: https://nces.ed.gov/fcsm/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf

[6] Information Quality Act. Pub. L. No. 106-554, § 515(a), 2000.

[7] OMB M-19-15. 2019. *Improving Implementation of the Information Quality Act.* Available at: https://www.whitehouse.gov/wp-content/uploads/2019/04/M-19-15.pdf

certainly requires this higher quality standard because of the universal impacts of the pandemic and the use of this data to inform decision making.

Figure 1. Federal Committee on Statistical Methodology *Framework for Data Quality*.



Source: FCSM-20-04. *A Framework for Data Quality.* Federal Committee on Statistical Methodology. Sept. 2020.

For the purposes of this analysis, we examined data quality across each of the three domains in this framework, with specific emphasis on six dimensions, which the research team of data experts deemed to be most relevant to the COVID-19 data examined: granularity, accessibility, timeliness, coherence, accuracy and reliability, and credibility (Table 1). The dimensions examined in detail are those most relevant for users of these data for both personal and policy decision-making related to COVID-19.

This assessment focuses on data available from the Federal government, which are reported by and aggregated from multiple actors including labs, hospitals, and state and local governments. The data examined were selected from the key indicators used and shared by HHS for the COVID-19 response, which include the volume of: testing, cases, hospitalizations, deaths, and vaccinations. We identified the primary sources of each of these indicators and the platforms and publications through which HHS shares these data with the public. Applying the domains and selected dimensions from the FCSM framework, we used these criteria to evaluate the quality of the data available for each indicator based on the information that is publicly available from HHS sources. Data related to COVID-19 have continued to evolve over the course of the pandemic; these findings reflect the data quality at the time this analysis was conducted, in 2021 and early 2022.

Table 1. Definitions of six dimensions of data quality examined.

Data Quality Measures		Definition
Domain	**Dimension**	
Utility	Accessibility	Accessibility relates to the ease with which data users can obtain an agency's products and documentation in forms and formats that are understandable to data users.
	Timeliness	Timeliness is the length of time between the event or phenomenon the data describe and their availability.
	Granularity	Granularity refers to the amount of disaggregation available for key data elements. Granularity can be expressed in units of time, level of geographic detail available, or the amount of detail available on any of a number of characteristics (e.g. demographic, socioeconomic).
Objectivity	Accuracy & Reliability	Accuracy measures the closeness of an estimate from a data product to its true value. Reliability, a related concept, characterizes the consistency of results when the same phenomenon is measured or estimated more than once under similar conditions.
	Coherence	Coherence is defined as the ability of the data product to maintain common definitions, classification, and methodological processes, to align with external statistical standards, and to maintain consistency and comparability with other relevant data.
Integrity	Credibility	Credibility characterizes the confidence that users place in data products based simply on the qualifications and past performance of the data producer.

Source: Definitions taken from FCSM-20-04. *A Framework for Data Quality.* Federal Committee on Statistical Methodology. Sept. 2020.

II.1. COVID-19 indicators and data sources

The data sources and flow of information for each of the COVID-19 indicators reviewed for this analysis are described below. The data sources for each indicator are summarized in Table 1. For each indicator, we focus on the total volume, which is a basic metric that can be converted into different rates using a unit of population, e.g., 100,000 people or the population of a defined geographic area, and a unit of time, e.g., day or week, and other derived metrics such as percent positivity from testing data.

Table 2. Summary of government and non-government data sources reviewed for key COVID-19 indicators.

COVID-19 Indicator	Federal Government Sources	Non-Federal Sources (acquired through web-scraping of local government websites)
Testing	CDC (through COVID-19 Electronic Laboratory Reporting)	
Cases	CDC (through the National Notifiable Disease Surveillance System)	New York Times
		Johns Hopkins University Center for Systems Science & Engineering
		Applied Physics Laboratory Aggregated Case & Death Counts (Johns Hopkins)
		USAFacts
		Conference of State Bank Supervisors
Hospitalizations	HHS Teletracking (daily reporting from all hospitals)	
	CDC COVID-NET (hospital sample system, covering 10% of US population)	
	CDC National Healthcare Safety Network (discontinued July 2020)	
Deaths	CDC (daily reporting from state and local governments)	New York Times
	NCHS/NVSS (provisional and final mortality statistics from death certificates)	Johns Hopkins University Center for Systems Science & Engineering
		Applied Physics Laboratory Aggregated Case & Death Counts (Johns Hopkins)
		USAFacts
		Conference of State Bank Supervisors
Vaccinations	CDC (IIS via IZ Gateway or CDC Data Clearinghouse)	

Testing

COVID-19 testing in the U.S. was developed by CDC in February 2020,[8] however, there were challenges regarding the accuracy of early tests. In response, FDA issued policy changes in February and March 2020, aimed to increase testing capacity by providing greater flexibility for laboratories and manufacturers of commercial test kits.[9]

The most reliable COVID-19 test is the laboratory-based Nucleic Acid Amplification Test (NAAT), also commonly referred to as a polymerase chain reaction (PCR) test or a molecular test.[10] Less accurate point-of-care NAATs and antigen tests are also available, which generally produce more rapid results. Point-of-care tests administered in settings such as schools and jails may be reported, but at-home tests are generally not captured in official reporting.[11] In many cases, a rapid test is confirmed with a follow-up laboratory test for a

[8] https://www.cdc.gov/coronavirus/2019-ncov/lab/testing.html
[9] https://www.federalregister.gov/documents/2020/05/15/2020-10492/policy-for-coronavirus-disease-2019-tests-during-the-public-health-emergency-immediately-in-effect
[10] https://www.cdc.gov/coronavirus/2019-ncov/lab/naats.html
[11] https://coronavirus.jhu.edu/pandemic-data-initiative/data-outlook/important-antigen-testing-data-are-inconsistent-and-unclear

more definitive diagnosis. Testing availability was limited early in the pandemic, and for several months, testing was more likely to be administered to symptomatic individuals, healthcare workers, and others who were considered at risk for exposure. In addition, testing data varied over time in terms of what information was reported, by whom, and how quickly.

The CARES Act, passed in March 2020, required all laboratories to report to the Secretary of HHS on COVID-19 tests, giving the Secretary the authority to stipulate required data elements.[12] To improve the consistency of data, CDC posted guidance for clinical laboratories reporting to states on May 6, 2020.[13] On June 4, 2020, HHS provided guidance for laboratories on additional data elements required when reporting COVID-19 test results, beginning in August 2020.[14]

Testing data have been reported to HHS through two major channels. The first, is the COVID-19 Electronic Laboratory Reporting (CELR), used by an increasing number of states as the pandemic progressed and to which all jurisdictions had converted as of April 2021.[15] Using this system, private, hospital, and public labs and state health departments report test results directly to CDC. The second (now outdated) channel, federal direct reporting, encompassed other systems by which data were reported without using CELR. This included large private labs that sent data directly to CDC; public health labs that reported data to CDC using the Public Health Laboratory Interoperability Project (PHLIP); and in-house hospital labs that reported data to HHS Protect, the internal HHS COVID-19 data platform, which was established soon after the start of the pandemic to provide a centralized location for all COVID-19-related data from across HHS and the Federal government, as well as from states and other partners, with access limited to authorized users.

Cases

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 notifiable conditions in CDC's National Notifiable Diseases Surveillance System (NNDSS).[16] COVID-19 cases are identified by hospitals, laboratories, and health care providers, who perform tests, and send the data to state and local health departments, which then submit these data voluntarily to the NNDSS. NNDSS data are considered provisional and subject to change until they are reconciled and verified with state and territorial data providers to be the final official incidence counts.

In addition to Federal case data reported through NNDSS, several non-governmental organizations compile case counts from local-level information, most often, from numbers published on health department websites. HHS Protect ingests, on a daily basis, estimates from several sources of this compiled data: the New York Times (NYT), the Johns Hopkins Center for Systems Science and Engineering (JHU CSSE), the Applied Physics Laboratory Aggregated Cases and Deaths Counts (APL ACDC), USAFacts, and the Conference of State Bank Supervisors (CSBS), as well as from CDC's NNDSS. Since December 2020, APL ACDC has been the "preferred" source of case estimates utilized by HHS Protect.

---

[12] https://www.congress.gov/116/bills/hr748/BILLS-116hr748enr.pdf
[13] CDC *Guidance for COVID-19 Pandemic Response, Laboratory Data Reporting: CARES Act Section 18115*, May 6, 2020
[14] CDC *Guidance for COVID-19 Pandemic Response, Laboratory Data Reporting: CARES Act Section 18115,* June 4, 2020
[15] https://www.cdc.gov/coronavirus/2019-ncov/lab/electronic-reporting-map.html
[16] https://ndc.services.cdc.gov/case-definitions/coronavirus-disease-2019-2020-08-05/

Hospitalizations

Data on COVID-19-related hospitalizations come from two primary sources – one which is more comprehensive in its detail and one which is more comprehensive in its representativeness. The first, the COVID-19-Associated Hospitalization Surveillance Network (COVID-NET), is a sample system that collects data from over 250 hospitals in 14 states, including detailed demographic information on each hospitalized patient.[17] The system is designed to capture data from communities that are collectively similar to, but not necessarily exactly representative of, the national population and provides the most complete information on demographic trends in hospitalizations.

As to the second source of hospitalization data, at the start of the pandemic, all hospitals reported data to CDC through the pre-existing National Healthcare Safety Network (NHSN), but this was replaced in July 2020 with a new system for hospital reporting directly to HHS.[18] This system, Teletracking, encompasses a set of indicators for which hospitals must submit data, either directly to HHS Protect or to states that then provide the data to HHS Protect. Reporting of the hospital Teletracking indicators has increased over time, particularly after CMS made this a condition of participation in the Medicare program, in August 2020.[19] As a result, nearly all hospitals report a set of required indicators to HHS on a daily basis – these include admissions, bed capacity, and supplies of necessary equipment, but patient breakdowns are collected only by age group with no other demographic characteristics.[20]

Deaths

Final mortality data for the United States are released annually, only after NCHS has received all death records from states and fully reviewed the data for completeness and quality. Final data contain the most accurate and complete information on mortality in the US**.** In response to the pandemic, and in order to provide timely data to help monitor pandemic-related mortality, the National Center for Health Statistics (NCHS) has been releasing provisional mortality data on COVID-19 deaths on a daily and weekly basis.[21] These data come directly from death certificates filed at the state and local level, and feature counts of COVID-19-related deaths by age, gender, race and ethnicity, place of death, and include information on other health conditions and comorbidities involved in these deaths. The provisional counts for COVID-19 deaths are based on the current flow of mortality data in the National Vital Statistics System (NVSS). National provisional counts include deaths occurring within the 50 states and the District of Columbia that have been received and coded as of the date specified.

In addition to the provisional and final mortality data collected and published by NCHS, states and jurisdictions report deaths on a daily basis to CDC, which CDC then validates through a confirmation process with each jurisdiction, in a similar manner to validation of case reporting. Also, in the same way that case reporting occurs through both government and non-government channels, non-government death counts are compiled from local sources by several organizations. HHS Protect ingests these death estimates on a daily basis from the same five non-government sources it relies on for daily case estimates: the New York Times, the Johns Hopkins Center for Systems Science and Engineering, the Applied Physics Laboratory Aggregated Cases and Deaths Counts (APL ACDC), USAFacts, and the Conference of State Bank Supervisors, as well as data reported daily to CDC. As with the case data, HHS Protect uses APL ACDC as the “preferred” source of death estimates.

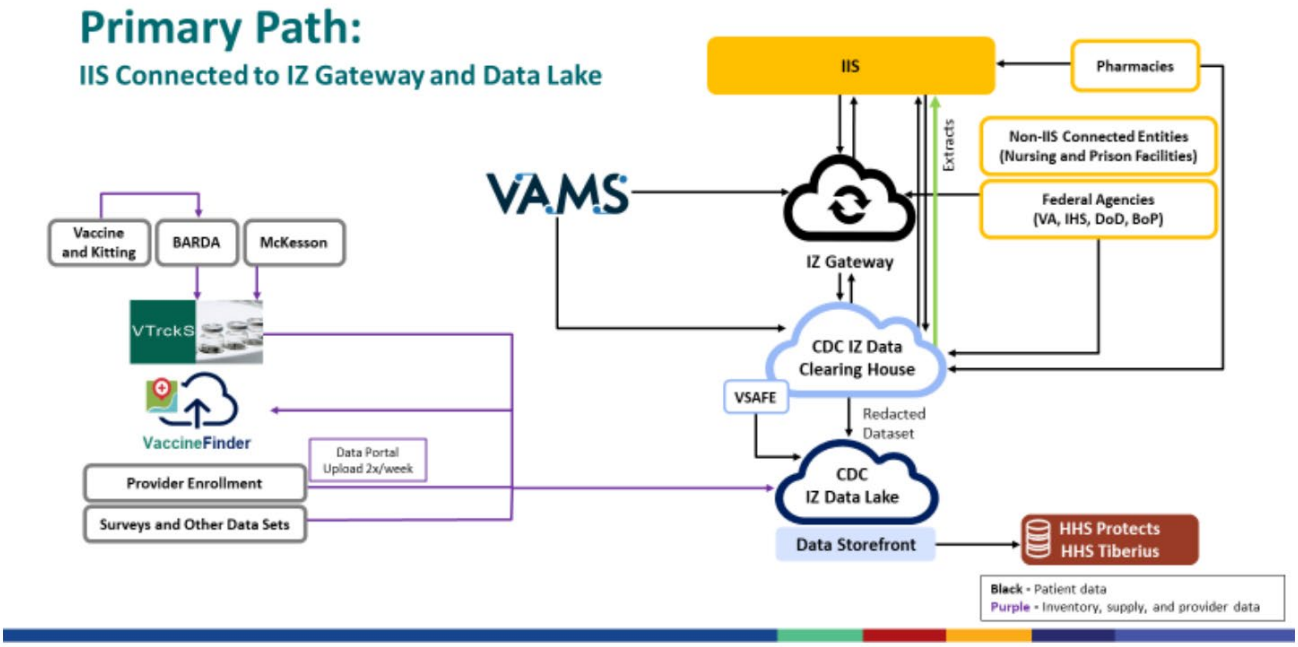
---

[17] https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covid-net/purpose-methods.html
[18] https://www.teletracking.com/news/hhs-renews-contract-with-teletracking-hhs-protect/
[19] https://www.cms.gov/files/document/covid-ifc-3-8-25-20.pdf
[20] https://www.hhs.gov/sites/default/files/covid-19-faqs-hospitals-hospital-laboratory-acute-care-facility-data-reporting.pdf
[21] https://www.cdc.gov/nchs/covid19/mortality-overview.htm

Vaccinations

CDC uses a combination of new and existing information technology systems to quickly collect data about COVID-19 vaccine administration and distribution (Figure 2). The Vaccine Tracking System (VTrckS) is CDC's vaccine order management system, which supports routine vaccinations with almost 80 million doses of vaccines annually.[22] CDC uses VTrckS as its platform for ordering all COVID-19 vaccines. VTrckS users are the 64 state, local, and territorial public health jurisdictions and enrolled national provider organizations (i.e., the Department of Veterans Affairs, Department of Defense, Indian Health Service, Bureau of Prisons, and pharmacy chains participating in COVID-19 vaccination).

Figure 2. Vaccine Administration Technical Architecture Diagram



Source: CDC Data Use and Sharing Agreement[23]

CDC's Vaccine Administration Management System (VAMS) is an online tool to manage vaccine administration from the time the vaccine arrives at a clinic through administration.[24] The Immunization Information Systems (IISs) vary by jurisdiction; state systems have varying capacity to automate processes and to handle a large volume of data. There can also be variations in data quality across IISs, sometimes due to data reporting policies.[25]

CDC's Immunization Data Clearing House is a cloud-hosted data repository that receives, de-duplicates, and de-identifies COVID-19 vaccination data that are then used to populate the IZ Data Lake with de-identified data for analytics. The Immunization Data Clearing House allows healthcare providers to search for a patient, see what brand of COVID-19 vaccine they received, and see when they received their first dose of COVID-19

---

[22] https://www.cdc.gov/vaccines/programs/vtrcks/index.html
[23] CDC. Data Use and Sharing Agreement to Support the U.S. Government's COVID-19 Emergency Response, Jurisdiction Immunization and Vaccine Administration Data Agreement. Available at: https://www.cdc.gov/vaccines/covid-19/reporting/downloads/vaccine-administration-data-agreement.pdf
[24] https://guest.vams.cdc.gov/?lang=en
[25] CDC. 2020. *Immunization Information Systems (IIS) Data Quality Blueprint*. Available at: https://www.cdc.gov/vaccines/programs/iis/downloads/Data-Quality-Blueprint-508.pdf

vaccine to ensure dose matching and appropriate vaccination intervals to complete the vaccine series. CDC, jurisdictions, federal agencies, and pharmacy partners use the IZ Data Lake to store and process administration, coverage, logistics, inventory, ordering, distribution, and provider data.

II.2. HHS public platforms for disseminating COVID-19 data

HHS and its component agencies have created several public data platforms and information sources to share the COVID-19 indicators discussed above with the public. However, the source data used to build these public tools vary, potentially contributing to the development of conflicting estimates.

The **CDC COVID Data Tracker** is the most comprehensive HHS public platform, providing a range of downloadable datasets on testing, cases, hospitalizations, deaths, and vaccinations, as well as emerging information on variants.[26] Many indicators are available at the county level and some data are provided with breakdowns by key demographics, including age, sex, race, and ethnicity, as well as for subpopulations such as residents of nursing homes and individuals who are incarcerated.

The **Community Profile Reports**, another HHS product, are produced by an interagency team coordinated by the White House COVID-19 Team.[27] These reports began in December 2020, were updated six days a week, and switched to twice a week in June 2021. They are distributed on HHS's open data site, **Healthdata.gov**. Each report provides multiple visualizations on testing, cases, hospitalizations, deaths, and vaccinations at the county level, as well as highlights localities with the greatest recent increases in cases. A county-level dataset is downloadable for each day these reports have been produced, starting in December 2020.

The **HHS Protect Public Data Hub** is the public-facing component of HHS Protect.[28] The HHS Protect Public Data Hub shares selected information from HHS Protect while protecting privacy and confidentiality, given the large volume of detailed and personal information compiled in HHS Protect. The focus of this site is hospital capacity, with detailed indicators on hospital utilization at the state and facility level, as well as county-level metrics for testing, cases, and deaths, and limited information about the distribution of therapeutics.

The National Center for Health Statistics provides provisional data from the National Vital Statistics System through its **COVID-19 Mortality Overview**.[29] Estimates of excess deaths, which may indicate additional mortality burden attributable to the pandemic, are also reported.

III. FINDINGS

This analysis assessed five key indicators of the COVID-19 pandemic – testing, cases, hospitalizations, deaths, and vaccinations – across all three domains of the FCSM data quality framework and for six specific dimensions within these domains. Table 2 provides a high-level summary of the level of data quality for each COVID-19 indicator by each dimension of quality examined.

[26] https://covid.cdc.gov/covid-data-tracker/#datatracker-home
[27] https://healthdata.gov/Health/COVID-19-Community-Profile-Report/gqxm-d9w9
[28] https://protect-public.hhs.gov/
[29] https://www.cdc.gov/nchs/covid19/mortality-overview.htm

Table 3. Summary of Findings: Degree to which each COVID-19 indicator meets the criteria for the assessed dimensions of data quality.

Data Quality Measures		COVID-19 Indicator				
Domain	**Dimension**	Testing	Cases	Hospitalizations	Deaths	Vaccinations
Utility	Accessibility	*	*	*	*	*
	Timeliness	*	*	*	^	*
	Granularity	!	!	^	*	!
Objectivity	Accuracy & Reliability	!	^	*	*	^
	Coherence	^	^	*	^	^
Integrity	Credibility	^	^	^	*	!

Green (*) = meets criteria; Yellow (^) = partially meets criteria; Red (!) = does not meet criteria

III.1. Data Quality by Domain

III.1.a. Utility

In the FCSM framework, utility is defined as: "the extent to which information is well-targeted to identified and anticipated needs. It reflects the usefulness of the information to the intended users."[30] Under this domain, we focused on data quality issues related to the dimensions of accessibility, timeliness, and granularity. Each dimension is defined and discussed in relation to each COVID-19 indicator below.

Accessibility

Accessibility "refers to the ease with which data users can obtain an agency's products and documentation in forms and formats that are understandable to data users."[31] Equitable access entails broad dissemination and ensuring transparency by providing complete documentation of dissemination policies. Data are considered accessible when they are available to a wide range of users in easy-to-understand formats and when metadata and documentation are provided to facilitate use and interpretation of the data. Accessibility also considers cost, access to proprietary information, accessibility for people with disabilities, and whether documentation is written in plain language. For this analysis, we examined the accessibility of data and documentation in terms of public availability only.

For the dimension of accessibility, all five indicators were determined to meet the criteria for data quality.

Accessibility by indicator:

Testing: Testing data are publicly available from healthdata.gov through a link from HHS Protect Public; the Community Profile Reports; and the CDC COVID-19 Data Tracker. The webpages where these data can be downloaded include descriptive text and variable definitions.

Cases: Estimated daily COVID-19 cases are published in the Community Profile Reports, on the CDC COVID-19 Data Tracker, and healthdata.gov. The CDC surveillance dataset is available in 12-element and 19-element public forms and a 32-element restricted form, for which users must request access. Variable

[30] FCSM. *A Framework for Data Quality*, pg. 21.
[31] FCSM. *A Framework for Data Quality*, pg. 22.

definitions are provided on the dataset webpage and data users are directed to the CDC case report form for detailed definitions of the data elements.[32]

Hospitalizations: The number of patients hospitalized with COVID-19 are available via facility-level datasets published on healthdata.gov from a link on HHS Protect Public. The complete Teletracking dataset of hospital indicators is not public, but most indicators are provided. The variables are defined on the webpage where data can be downloaded. A more detailed form with instructions to hospitals for completing each data element is provided on a separate HHS webpage, not linked from the dataset page.[33] The COVID-NET hospital surveillance dataset can be downloaded from the CDC website; descriptive text is provided.

Deaths: Estimated daily COVID-19 deaths can be downloaded from the CDC COVID-19 Data Tracker, are published in the Community Profile Reports, and provisional COVID-19 deaths can be downloaded from the NCHS COVID-19 Mortality Overview. NCHS also publishes excess mortality associated with COVID-19. The webpage where NCHS data can be downloaded contain descriptive text, including information on: comparing the data to other sources; the nature and sources of the data; cause-of-death classification and definition of deaths; estimated completeness of data; delays in reporting; estimated distributions of COVID-19 deaths and population size by race and Hispanic origin; quality of race and Hispanic origin data; estimates of disparities and adjusting for age; place of death; and, comparing deaths from different states.

Vaccinations: COVID-19 vaccination data are available from the CDC COVID-19 Data Tracker and the Community Profile Reports. Technical notes and variable definitions are provided on the webpage where data are available.

Timeliness

Timeliness reflects the length of time between an event described by data and the availability of that data. There is generally a trade-off between timeliness and other domains of data quality, including accuracy and granularity.

For the dimension of timeliness, four of five indicators were determined to meet the criteria for data quality; deaths were found to partially meet the criteria.

Timeliness by indicator:

Testing: Jurisdictions report testing data to CDC on a daily basis but may be delayed up to several days to mitigate discrepancies due to reporting variations. The testing dataset linked from HHS Protect Public is updated each weekday for the preceding day. Testing data on the CDC COVID-19 Data Tracker are updated each day except Sunday.

Cases and Deaths: The timeliness of COVID-19 case and death data can broadly be distinguished in two groups: fairly immediate provisional data and more delayed final data. This distinction separates the case and death data that are compiled by many organizations on a daily basis from local government sources (such as websites) versus data sent to CDC. Many jurisdictions send case reports to CDC daily, but some report less often; provisional death estimates are reported to NCHS with a delay of 1-2 days, and final death certificate data are reported to CDC with about an 8-week delay. Case and death data on the COVID-19 Data Tracker are updated six days a week.

[32] https://www.cdc.gov/coronavirus/2019-ncov/downloads/pui-form.pdf
[33] https://www.hhs.gov/sites/default/files/covid-19-faqs-hospitals-hospital-laboratory-acute-care-facility-data-reporting.pdf

Hospitalizations: The timeliness of data on COVID-19-related hospitalizations greatly improved over the course of the pandemic. Since reporting to the NHSN was replaced with the Teletracking system and participation became mandatory when CMS made this a requirement, nearly all hospitals have reported this information on a daily basis. Facility-level hospitalization data are available via HHS Protect Public on a weekly basis. The COVID-NET data on the CDC COVID Data Tracker are also updated weekly.

Vaccinations: Jurisdictions are required to report vaccination data to CDC twice per week and have a reporting lag of up to five days. Vaccination data are updated on the CDC COVID Data Tracker daily.

Granularity

Granularity refers to the amount of disaggregation available for key data elements and has been an acknowledged weakness in the quality of COVID-19 data. Key aspects for which granularity is relevant to COVID-19 include geographic location and patient demographics such as age, sex, race, and ethnicity, as well as other personal attributes such as type of employment, income level, underlying medical conditions, and type of residence. Details on each of these characteristics help to reveal the extent to which the pandemic is differentially affecting specific subpopulations. The need for hyper-local information on the pandemic, for example, publicly reporting cases by census tract, is important, but COVID-19 data have not been widely available at a more refined geographic level than counties.

For the dimension of granularity, death data were determined to meet the criteria for data quality, hospitalizations were found to partially meet the criteria, and testing, cases, and vaccinations were found to not meet the criteria for data quality.

Granularity by indicator:

Testing: Most testing data are reported to HHS without patient demographic information. This information is not federally required to be collected when a test is given.

Cases: The official case report form utilized by CDC has fields for county, age, sex, race, ethnicity, hospitalization status, type of residence, health worker status, and underlying conditions, but these are not uniformly completed.

Hospitalizations: There are estimates of the racial and ethnic distribution of COVID-19-related hospitalizations from COVID-NET, the hospital sample system, but this does not capture all hospitalized cases. Due to the shifting dynamics of the pandemic over time, the representativeness of patients in hospitals participating in this sample may not always accurately reflect the national context. Hospitalization data collected by Teletracking include patients by age group but do not include other demographics.

Deaths: Deaths are the one type of key COVID-19 data for which more comprehensive demographic information is consistently and widely available. Official death information is reported using a federal death certificate, which includes data fields for state of residence, education, marital status, age, sex, race and ethnicity. Race and ethnicity are collected in accordance with OMB and HHS standards.

Vaccinations: Most vaccination data are reported to HHS without patient demographic information, aside from sex, date of birth, and address. Information on race and ethnicity are not federally required to be collected when a vaccine is given.[34]

III.1.b. Objectivity

Objectivity is defined by FCSM as "whether information is accurate, reliable, and unbiased, and is presented in an accurate, clear, and unbiased manner."[35] Under this domain, we examined data quality issues related to the dimensions of accuracy and reliability as well as coherence. Each dimension is defined and discussed in relation to each COVID-19 indicator below.

Accuracy and Reliability

Accuracy describes how closely a number or estimate from a data source reflects its true value, and reliability describes the consistency of results when reproduced or measured under similar conditions. A significant challenge related to determining the accuracy and reliability of COVID-19 data is that the data originate from numerous and different sources, including reporting from labs, providers, and jurisdictions. To truly determine accuracy and reliability requires identification and assessment of each of the individual sources that are aggregated to create national and subnational metrics. A key aspect of accuracy and reliability that is more feasible to assess at the national level is the level of completeness or the proportion of missing values, which is also a focus of the discussion below.

For the dimension of accuracy and reliability, testing data were determined not to meet the criteria for data quality, cases and vaccinations were found to partially meet the criteria, and hospitalizations and deaths were found to meet the criteria for data quality.

Accuracy and reliability by indicator:

Testing: An important dimension of test data accuracy is diagnostic test performance. In clinical settings, the PCR test for COVID-19 has a sensitivity of almost 80% and over 98% specificity;[36] however, rapid tests have lower sensitivity and specificity.[37] The accuracy of testing data may be declining as increasing use of (less accurate) rapid tests replaces (more accurate) laboratory tests. Estimates of the completeness of patient demographic information in COVID-19 testing data are not publicly available.

Cases: The accuracy of case data was initially impacted by the absence of data on nursing homes early in the pandemic because case reporting from nursing homes was optional before May 8, 2020. As noted above, the switch to more rapid tests, particularly those administered at home, which are largely unreported, and a reduction in contract tracing may lead to a decline in the accuracy of case counts. As of March 2022, the completeness of patient demographics in the CDC public case dataset was: 74.4% for race, 68.6% for ethnicity, 98.5% for sex, and 98.9% for age group. County of residence was available for all cases.[38]

---

[34] https://www.cdc.gov/vaccines/covid-19/vaccination-provider-support.html
[35] FCSM. *A Framework for Data Quality*, pg. 23.
[36] https://www.cap.org/member-resources/articles/how-good-are-covid-19-sars-cov-2-diagnostic-pcr-tests
[37] https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antigen-tests-guidelines.html#table1
[38] https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4. Authors' calculations of data available on March 9, 2022. Response categories "unknown" and "missing" were counted as incomplete. Suppressed cells were omitted.

Hospitalizations: The accuracy of the count of COVID-19-related hospitalizations is likely fairly high since hospitals began to be required to report this information in August 2020, however misclassification of patients by the cause of hospital admission can occur. The completeness of race and ethnicity in hospitalization data from COVID-NET is high, but specific estimates of completeness by indicator are not publicly available.

Deaths: While there is sometimes misclassification of causes of death, national mortality data are generally highly accurate and reliable. Provisional COVID-19 death data from the National Vital Statistics System have high completeness for all patient demographics. Age is complete, sex is more than 99.9% complete, and race and ethnicity are 99% complete. Assessments of the validity of reporting race and ethnicity and death certificates have found record-level agreement between death certificates and Census data close to 100% for both the White and Black populations. Agreement for Hispanic populations was 90%. Agreement on the identification for American Indian and Alaska Natives was lower at about 60%.[39]

Vaccinations: The accuracy and reliability of vaccination data are hampered by notable gaps in four states' data compiled by CDC: Hawaii does not provide the county of residence for vaccine recipients, Texas provides state-level data only (nothing at the county level), California does not report county of residence for counties with populations below 20,000, and Massachusetts does not provide data for three low-population counties.[40] The completeness of demographics in vaccination data is currently highly uneven.

Coherence

Coherence is defined as how well a data product reflects common definitions, classifications, and methodological processes, aligns with external statistical standards, and maintains consistency and comparability with other relevant data. Coherence applies to data over time, across key domains and when data originate from different internal and external sources.[41]

For the dimension of coherence, hospitalization data were determined to meet the criteria for data quality; the other four indicators were found to partially meet the criteria.

Coherence by indicator:

Testing: The vast majority of testing data, particularly for tests that are reported, come from one of several approved brands of laboratory NAATs, which produce comparable results.[42] At-home and point-of-care rapid tests provide less consistent results and are also not systematically reported.

Cases: As shown in Figure 3, there is variation across sources of COVID-19 case estimates. These variations are due to one or more factors: the inclusion (or not) of probable (not confirmed) cases, how non-residents in a state are counted, cases not attributed to a specific county (unallocated), and the inclusion or exclusion of incarcerated populations. As is evident in comparing data points for the most recent months versus earlier months, coherence improves over time – likely due to the investigation and resolution of discrepancies.

---

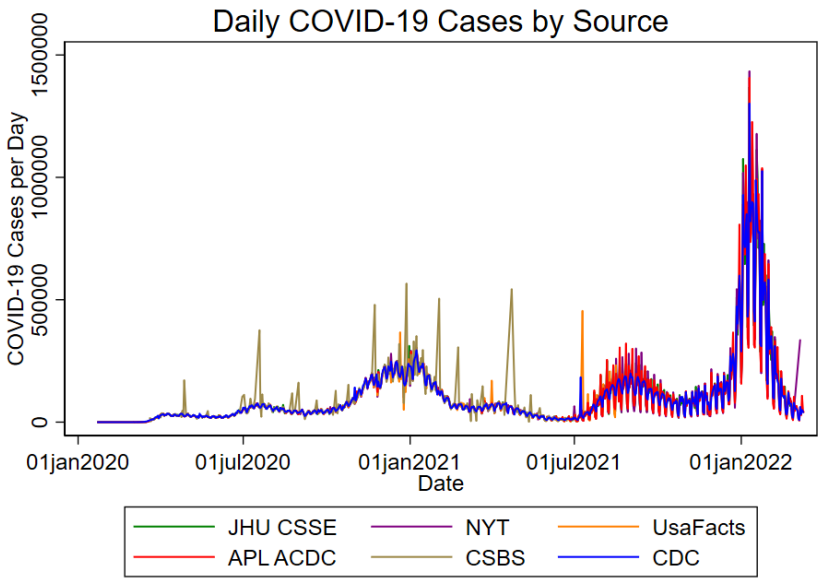
[39] Arias E, Heron M, Hakes JK. *The validity of race and Hispanic-origin reporting on death certificates in the United States: An update.* National Center for Health Statistics. Vital Health Stat 2(172). 2016

[40] https://www.cdc.gov/coronavirus/2019-ncov/vaccines/distributing/reporting-counties.html

[41] FCSM. *A Framework for Data Quality*, pg. 24.

[42] https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/in-vitro-diagnostics-euas-molecular-diagnostic-tests-sars-cov-2

Figure 3. Coherence of daily COVID-19 case estimates across sources, for all sources compiled by HHS Protect.



Source: Authors' analysis of HHS Protect data

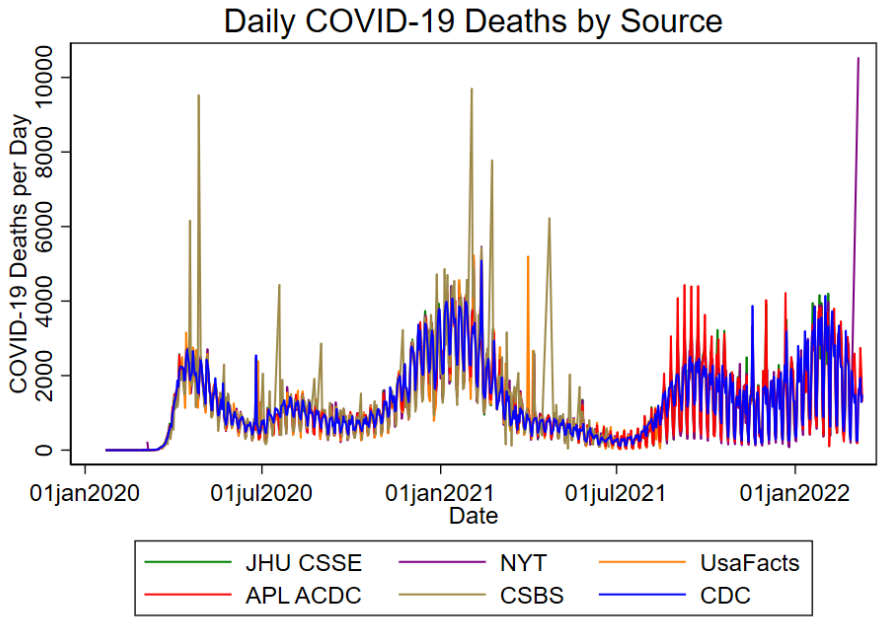
Hospitalizations: COVID-NET data collection is highly standardized: surveillance officers compile data from medical records using standard case reporting. Teletracking indicators are also reported using standardized definitions, although shifting reporting requirements over the course of the pandemic have caused disruptions as hospitals adjust to new requirements.

Deaths: Final mortality data from the National Center for Health Statistics (NCHS) contain the most coherent and standardized information on national mortality. NCHS compiles mortality statistics in accordance with World Health Organization (WHO) regulations to classify and code causes of death in alignment with the current revision of the *International Statistical Classification of Diseases and Related Health Problems* (ICD),[43] including the new code (U07.1) for deaths due to COVID-19.[44] NCHS data may differ slightly from other sources due to differences in completeness, COVID-19 definitions used, and delays in reporting. Provisional death estimates from sources other than NCHS (Figure 4) vary from one another and are subject to the same limitations as case data: variation across sources due to the inclusion (or not) of probable (not confirmed) COVID-19 deaths, how non-residents in a state are counted, deaths not attributed to a specific county (unallocated), inclusion or exclusion of incarcerated populations, or other factors.

[43] https://www.who.int/standards/classifications/classification-of-diseases

[44] National Vital Statistics System. 2020 *Guidance for Certifying Deaths Due to Coronavirus Disease 2019 (COVID-19).* Vital Statistics Reporting Guidance Report No. 3. Available at: https://www.cdc.gov/nchs/data/nvss/vsrg/vsrg03-508.pdf

Figure 4. Coherence of daily COVID-19 mortality estimates across sources, for all sources compiled by HHS Protect.



Source: Authors' analysis of HHS Protect data

Vaccinations: According to CDC, vaccination data reported by CDC can differ from estimates reported on state and local websites due to "differences in how data were reported or how the metrics are calculated."[45]

III.1.c. Integrity

Integrity is defined by FCSM as "the maintenance of rigorous scientific standards and the protection of information from manipulation or influence as well as unauthorized access or revision." Under this domain, we examined data quality issues related to the dimension of credibility. This dimension is defined and discussed in relation to each COVID-19 indicator below.

Credibility

Credibility "characterizes the confidence that users place in data products based simply on the qualifications and past performance of the data producer."[46] Transparency, a closely related concept, is a component or practice that provides credibility by ensuring information needed to understand the quality and limitations of the data is provided. Credibility is increased when the methods and approaches used to collect, process, and analyze the data, and the reproducibility of findings from the data, are transparent.

Transparency is particularly important for conveying data quality when there is more than one source in a data product[47] – a single data source has threats to quality, but with integrated data products, such as combining counts of tests or deaths from different states, data quality must be assessed for all source data as well as the

[45] https://www.cdc.gov/coronavirus/2019-ncov/vaccines/distributing/reporting-counties.html
[46] FCSM. *A Framework for Data Quality*, pg. 26.
[47] FCSM. 2018. *Transparent Quality Reporting in the Integration of Multiple Data Sources: A Progress Report, 2017-2018*. Federal Committee on Statistical Methodology. October 2018. Available at https://nces.ed.gov/fcsm/pdf/Quality_Integrated_Data.pdf.

data products that result.[48] The transparency of information provided on data sources, data collection methods, data processing, limitations and data quality varies considerably depending upon both the data source and the platform from which data are being provided.

For the dimension of credibility, death data were determined to meet the criteria for data quality; vaccinations did not meet the criteria; the other three indicators were found to partially meet the criteria for data quality.

Credibility by indicator:

Testing: The source cited for the testing data reported in the Community Profile Reports is the "unified testing dataset," but no details are provided regarding how this dataset is compiled, aside from describing the CELR data reported by states as the primary source of information and data reported directly by labs when the other data are not available. A list of states and counties reporting via CELR is provided. The testing data presented on the CDC COVID Data Tracker provide similar details in the footnotes. Likewise, similar source information is provided on the testing data page of HHS Protect Public, but the national-level metrics on the homepage, including testing, provide no information on data sources.

Cases: The case estimates presented in the Community Profile Reports list CDC as the source of data and provide an abbreviated version of the source details that are available in the footnotes of the case data presented on the CDC COVID Data Tracker. However, the utilization of other sources, which are also compiled by HHS Protect and inform the estimates released by HHS, is not acknowledged. The national-level metrics on the homepage of HHS Protect Public, including cases, provide no information on data sources.

Hospitalizations: The hospitalization data presented in the Community Profile Reports cite the "Unified Hospital Dataset" as the source but provide only limited details regarding how this dataset is compiled. This dataset is also cited as the source for the hospitalization data on the CDC COVID Data Tracker, but no details are provided aside from this data being "based on reporting from all hospitals." The hospitalization data from COVID-NET presented on the CDC COVID Data Tracker provide more detailed source information. The detailed hospitalization data on HHS Protect Public links to a healthdata.gov page with source information, but the national-level metrics on the HHS Protect Public homepage, including hospitalizations, provide no information on data sources.

Deaths: The COVID-19 mortality estimates presented in the Community Profile Reports cite CDC as the data source and, similar to case data, do not acknowledge the integration of non-government sources in these estimates; for these data, similar information to the footnotes on the CDC COVID Data Tracker are provided. The mortality data from the NCHS COVID Mortality Overview include detailed technical notes on provisional, final, and excess mortality. The national-level metrics on the homepage of HHS Protect Public, including deaths, provide no information on data sources.

Vaccinations: The vaccination data presented in the Community Profile Reports cite the source as the "Unified COVID-19 Vaccine Dataset" but provide no further details on this data source. The vaccination data available from the CDC COVID Data Tracker include descriptive footnotes and a link to a separate page with additional details.

[48] Czajka JL, Stange M. 2018. Transparency in the Reporting of Quality for Integrated Data: A Review of International Stan-dards and Guidelines. Washington, DC: Mathematica Policy Research, April 27, 2018. Available at: https://www.mathematica.org/our-publications-and-findings/publications/transparency-in-the-reporting-of-quality-for-integrated-data-a-review-of-international-standards

IV. DISCUSSION

Across the data quality domains of utility, objectivity, and integrity, the dimensions of timeliness, granularity, accessibility, accuracy and reliability, coherence, and credibility provide important attributes to assess data quality. Key data used to characterize and guide the response to the COVID-19 pandemic include the volume of testing, cases, hospitalizations, deaths, and vaccinations. Examining these indicators in relation to these data quality criteria reveals limitations to consider when using the data; characteristics that make each indicator the most appropriate choice for certain purposes; and highlights the availability of information from different sources, which may in some cases conflict.

For tracking the progress and status of the pandemic, case and death data provide the best indication of the prevalence of COVID-19 over time, generally providing a reliable picture of disease spread. However, an important caveat is the low availability of testing early in the pandemic, leading to a likely underestimate of the volume of cases in the earliest stage. Cases are assumed to be undercounted throughout the pandemic, with many people being asymptomatic, choosing to forgo testing, or unable to access testing, but the proportion of unreported cases likely stabilized once the availability of testing became more consistent. The pattern of deaths also provides an indication of the relative prevalence over time, although the case fatality rate declined as treatment improved over the course of the pandemic; as it reached younger populations versus the initial peak among the elderly, for whom mortality rates are higher; and as vaccinations, which greatly reduce the chance of death, became widespread.

Excess mortality provides the best picture of the full mortality burden related to the COVID-19 pandemic, by including deaths that may be indirectly attributed to COVID-19. Excess deaths are typically defined as the difference between the observed number of deaths and expected number of deaths in the same time period, based on past trends. Excess deaths may include deaths due to COVID-19 that were not correctly attributed to the virus and also deaths related to delayed or foregone care due to the pandemic's impacts on the health system and deaths due to conditions exacerbated by economic hardship or other conditions related to the pandemic.

Given the critical need to understand disparities in the impacts of the pandemic, it is vital to measure differences in testing, cases, hospitalizations, deaths, and vaccinations by race and ethnicity. Mortality derived from death certificates is the only COVID-19 metric with near-complete reporting of race and ethnicity and should be considered the most reliable measure of racial and ethnic disparities in the impact of COVID-19. However, the case fatality rate of COVID-19 is likely not uniform across subpopulations due to differences in rates of underlying conditions, access to health care, and other factors; as a result, disparities in mortality may differ from disparities in cases. Hospitalization data from COVID-NET also provide near-complete race and ethnicity information but have the limitation of coming from a sample as opposed to capturing all hospitalizations nationwide. Data on cases, testing, and vaccinations should be considered less reliable for assessing disparities, given the consistent incompleteness of race and ethnicity information in these data.

Challenges

One of the primary data quality issues negatively impacting COVID-19 data is a lack of clarity and transparency regarding the most reliable source of information, given multiple estimates available from both government and non-government sources. This challenge is rooted in the multiple levels of data collection systems, from local to state to federal, and in the case of case and death data, several non-government organizations that compile data directly from local sources and produce alternative county, state, and national estimates. HHS uses these non-government sources of case and death data, but does not transparently report how these sources are integrated with CDC data for official reporting. For hospitalization data, an early source of confusion came from the choice to switch hospital reporting from CDC to HHS – ending NHSN and establishing

Teletracking as a new system. For testing data, the availability of at-home tests and the lack of visibility into all providers administering point-of-care COVID-19 tests mean there is incomplete documentation of all tests conducted and their results. Differences also arise between provisional and final values for most indicators, as data are shared on a frequent basis but continually updated.

This complexity of data sources underlies limitations in the accessibility of COVID-19 data and its documentation. While these data are widely available from HHS and are provided at no cost, the completeness and ease of comprehension of the data vary by source. Overall, a large volume of COVID-19 data is accessible online, but it requires some expertise and familiarity with the data to identify the official sources of different estimates and some detailed data remain unavailable. For example, NCHS mortality data files are free and available for download by the public, while the full Teletracking hospital dataset is not publicly available (selected indicators are withheld). Testing data are readily available via HHS Protect, but this excludes antigen testing, leaving the picture incomplete. CDC publishes vaccination totals on its website and has added downloadable public files, but selected counties are missing. The publication of Community Profile Reports and the release of facility-level hospital data, both starting in December 2020, were milestones in the provision of information to the public, but the lack of substantial details on data sources in the Community Profile Reports remains a limitation.

These challenges affect the credibility of federal COVID-19 data. An August 2021 GAO report noted credibility concerns with HHS's data: "Several stakeholders raised that HHS Protect and Teletracking may lack sufficient data quality checks".[49] However, it is possible that these concerns were based primarily on experiences with the data in the early days of these systems; procedures to validate and correct data were instituted after several months of operation. A key conclusion of this GAO report was the importance of dialogue with affected stakeholders, seeking input, and improving communication when instituting changes to reporting requirements. A similar recommendation to establish an expert committee was made by a January 2021 GAO report,[50] but was not implemented as of this second report. Transparency is crucial to credibility and can be improved with public consultation, clarity of source data, and the provision of complete documentation for all data.

An important limitation of testing, case, death, and vaccination data is the lack of granularity for information specific to geographic locations below the county level; however, hospitalization data are available at the facility level. Such granular data are needed to guide individuals' decision-making about the risk associated with different activities at different points in the pandemic and community-level decisions, such as when to reopen schools and whether to institute or repeal mask mandates. As the pandemic is in a stage characterized by substantial, but not universal, vaccine coverage, there is a need for hyper-local information on vaccination rates so individuals and policy officials can understand the level of risk in different locations, which will vary based on the number of unvaccinated people.

Although COVID-19 data come from a number of sources, we focused our detailed assessment of quality domains and dimensions on data available from the federal government. A number of data quality issues have also been identified by investigations and assessments of COVID-19 data reporting at the state and local levels. These concerns arise from some states failing to collect patient characteristics for the vast majority of cases, outdated technology for reporting at the county level, a lack of interoperability between reporting systems within states, and the absence of state authority to mandate data collection by localities, among other

[49] Government Accountability Office. *COVID-19: HHS's Collection of Hospital Capacity Data*. GAO-21-600. August 2021. https://www.gao.gov/assets/gao-21-600.pdf

[50] https://files.gao.gov/reports/GAO-21-265/index.html

issues.[51] Any data quality concerns in local-level data are carried forward into data compiled for federal reporting and should be considered as affecting the underlying data.

An important context for this assessment is the novelty of the data needs created by the COVID-19 pandemic, which required new and more comprehensive forms of public health data collection than have previously existed at the national level in the United States. Systems for COVID-19-related data collection were built and strengthened throughout the pandemic, leading to steady improvements in data quality over time. Arguably the most important lesson from these data gathering efforts is identifying necessary changes to be better prepared for a future pandemic or public health emergency.

Future Considerations

HHS has compiled and released an unprecedented volume of data during the COVID-19 pandemic, empowering the public to examine and analyze various aspects of its trajectory and impacts. This positive momentum towards making data more open, which the federal government has embraced as a priority, should be continued, but the quality of data must also be continually assessed and improved. Major steps HHS could take to improve the quality of data on COVID-19 testing, cases, hospitalizations, deaths, and vaccinations are outlined below and summarized in Table 4.

Table 4. Summary of recommendations, and corresponding data quality dimensions, to improve the quality of federal COVID-19 data.

Data Quality Measures		Recommendations
Domain	**Dimension**	
Utility	Accessibility	- Emulate best practices used by NCHS and approaches recommended by the Committee on National Statistics to develop and disseminate resources to guide the collection and reporting of COVID-19 data.
	Timeliness	- Correct data collection gaps encountered at the start of the pandemic to ensure accurate, timely, and granular data are available when needed. - Assess whether to maintain new data collection systems established during the pandemic
	Granularity	
Objectivity	Accuracy & Reliability	
	Coherence	- Strive for coherence across all HHS-published estimates and clearly indicate the reasons for any discrepancies. - Release comprehensive data standards that cover the most frequently reported COVID-19 metrics that all data collectors should follow.
Integrity	Credibility	- Increase the transparency of source information for all COVID-19 data.

First, HHS could increase the transparency of the source of information, which can build the credibility of HHS as a provider of COVID-19 data. In the HHS Protect Public Data Hub, the unified case and death datasets could be made available to the public, with an explanation of the sources compiled to create these data, including clarifying the role of non-government data in informing the estimates and the rationale for the preferred data series. The national metrics presented on this website should provide source information. The Community

[51] https://www.npr.org/2021/09/01/1032885251/millions-of-people-are-missing-from-cdc-covid-data-as-states-fail-to-report-case

Profile Reports could specify the data sources used in greater detail. The provision of clear and visible information on data sources should be a priority for all HHS publications of COVID-19 indicators.

To minimize confusion, HHS and its component agencies should strive for coherence across their published estimates and clearly indicate the reasons for any discrepancies. This includes actions like ensuring the HHS Protect public dashboard of national metrics is updated with the same frequency as the CDC COVID Data Tracker dashboard and adding highly visible language that differentiates between death data acquired daily by CDC from jurisdictions and provisional death certificate data collected by NCHS and presented on the NCHS COVID-19 Mortality Overview and the CDC COVID Data Tracker pages. Discrepancies between sources could also be reduced by the release of comprehensive data standards from HHS that cover the most frequently reported COVID-19 metrics that all data collectors should follow. This could include how probable cases should be counted, guidelines for assigning cases to locations when residency information is missing, procedures for counting institutionalized populations, and other factors that currently contribute to differences in estimates.

Several new data systems were established in response to COVID-19, filling existing gaps in the federal public health data architecture. This includes Teletracking, the daily data collection from all of the nation's hospitals, and HHS Protect, the internal HHS data hub for reporting and aggregating large volumes of COVID-19-relevant information. HHS must decide whether these systems should continue in order to be better prepared for a future pandemic, whether changes are needed to operate more effectively, or whether to retire these when the pandemic ends. If not continued or improved, the federal government will need to assess whether changes are needed to existing data collection systems to be prepared for a similar emergency in the future and correct gaps encountered at the start of this pandemic to ensure accurate, timely, and granular data are available when needed.

In striving to produce and provide high quality data to the public, NCHS represents best practices for other HHS agencies to emulate along with approaches recommended by the Committee on National Statistics.[52] To facilitate high quality COVID-19 mortality data, NCHS developed a number of resources for states and providers reporting deaths related to COVID-19. Examples of these resources include information on the new ICD-10 code for COVID-19; *Guidance for Certifying Deaths Due to COVID-19*;[53] reports assessing the timeliness of death certificate data; methodological reports assessing new methods to account for lags in reporting and identifying underlying causes of death that are unsuitable for assessing quality. These continual and routine evaluations are an important component for assessing the quality of data. The development and dissemination of similar resources to guide the collection and reporting of other types of COVID-19 data could help to improve data quality.

V. CONCLUSION

This analysis applied an established data quality framework to federal data on COVID-19 testing, cases, hospitalizations, deaths, and vaccinations across the data quality domains of utility, objectivity, and integrity, with a focus on the dimensions of timeliness, granularity, accessibility, accuracy and reliability, coherence, and credibility. The findings of this analysis represent a point in time and some conclusions may change given the dynamic nature of COVID-19-related data. Overall, mortality data from NCHS are the highest quality and

[52] CNSTAT. National Academies of Sciences, Engineering, and Medicine. 2019b. *Methods to Foster Transparency and Reproducibility of Federal Statistics: Proceedings of a Workshop.* Washington, DC: The National Academies Press. Available at: https://www.nap.edu/catalog/25305/methods-to-foster-transparency-and-reproducibility-of-federal-statistics-proceedings
[53] https://www.cdc.gov/nchs/data/nvss/vsrg/vsrg03-508.pdf

testing data are the lowest quality. Key challenges identified include the broad range of government and non-government data sources and the lack of coherence between different estimates; the inaccessibility of clear and complete source documentation; and reduced credibility arising from lack of transparency about data sources. Furthermore, documented weaknesses in local-level data collection and reporting reduce the quality of the inputs to federal data. With the exception of hospitalizations, which are available at the facility level, COVID-19 indicators are not available with a sufficient level of geographic granularity to effectively inform individual or local policy official decision-making. When utilizing COVID-19 data for different applications, case and death data provide the best indication of prevalence and the pandemic's progression over time; excess mortality is the best measure of the full extent of COVID-19's impacts; and NCHS mortality data provide the best measures of racial and ethnic disparities. The identified challenges are not unusual for new data collection and do not indicate that these data should not be used, but rather, that the data are best used with an understanding of their strengths and weaknesses. There remain steps that HHS is taking and can take to continue to improve the quality of COVID-19 data. Priority areas include increasing transparency through the provision of clear and visible information on data sources in all HHS publications of COVID-19 indicators; ensuring all HHS published COVID-19 data are coherent; deciding whether to maintain the new data systems established during the pandemic and, more broadly, planning to address the data collection gaps exposed by COVID-19 in preparation for future pandemics and health emergencies. Lastly, all HHS agencies producing and sharing data could emulate the practices of NCHS, the Department's statistical agency, in the dissemination of detailed information and documentation for data users.