ORNL/TM-2022/1496

DOE COVID-19 Data Curation Effort: Overview of Initial Data Collection Coverage (March – June 2020)

Jesse Piburn
Jason Kaufman
Alex Sorokine
Robert Stewart

August 2022

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE US DEPARTMENT OF ENERGY

ORNL/TM-2022/1496

National Security Emerging Technologies Division

DOE COVID-19 Data Curation Effort: Overview of Initial Data Collection Coverage (March – June 2020)

Jesse Piburn
Jason Kaufman
Alex Sorokine
Robert Stewart

Date Published:

August 2022

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-BATTELLE, LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

FIGURES

1. EXECUTIVE SUMMARY

During the COVID-19 pandemic of 2020, major case reporting outlets quickly coalesced around two or three primary vendors. Johns Hopkins University and *The New York Times* were among the more prominent, and all were of great value to the nation, particularly during the uncertain early stages of the pandemic. They primarily focused on three major attributes: number of new cases, deaths, and recovery, but only at the state level. Recognizing that many states were reporting very detailed data sets (e.g., hospital beds) at a count level or finer, the ORNL Pandemic Modeling team embarked on a major data curation effort from March to June 2020 for the purpose of capturing this wealth of detailed data. The challenge of curating this data was daunting. The number of attributes reported by the states grew on almost on a weekly basis. States were routinely shifting their web tool strategies away from easily parsable HTML-based formatting to new Tableau and ArcGIS content. This growth in the sheer number of attributes combined with the unpredictable shifts in data format meant an aggressive and agile combination of automated scripting and manual scraping was required to capture new daily streams. To keep up, the team had to scale up staff and widen its approach for capture and storage.

Ultimately, the collection of more than 11 million data points would not have been possible otherwise. By the end of Phase I, curation had largely ceased, giving way to new priorities in Phase II, during which the team turned its attention to cleaning up and harmonizing the data. In capturing the raw data, each state followed its own naming conventions, often evolving over time, causing direct analysis of the raw data to be impossible. A substantial, and ongoing, harmonization effort was required to merge and link these data. Also, categories of data, particularly demographic data, were broken out differently by different states. Missing data had to be tracked down, along with a wide variety of other challenges. In the end, a subset of the most common data was harmonized and stored in a PostGreSQL database where it could be easily accessed and analyzed. This report provides a top-level view of the inventory and completeness of this database. A follow-on report will focus on major trends and patterns in the data.

The DOE COVID-19 data collection effort resulted in over 11 million data points being collected, covering over 13,000 unique geographies and over 2,000 unique attributes that spanned predominantly from early March through the end of June 2020.

County-level reports of cases, deaths, and later in the summer, testing, broken down by demographic categories, such as age, sex, and race, provide the largest overlap to support deeper nationwide analysis. For more spatially specific analysis, such as within a certain state, more detailed analysis could be possible as a substantially long tail of location-specific attributes was collected as well.

1.1 KEY FINDINGS

- Over 11 million individual data points were collected by the ORNL COVID-19 curation effort
- Coverage includes data from all 50 states and multiple territories
- State- and county-level reporting are the most common, but more than 13,000 unique locations appear in the database, such as zip codes, health regions, and census blocks
- Most data points range from mid-March 2020 through the end of June 2020
- Types and number of attributes available vary widely across states
- Most states reported at the county level, with increasing demographic detail throughout the collection period; however, numerous other reporting geographies are present in the data, including health regions, zip codes, cities, census blocks, and in some cases, individual facilities (e.g., nursing homes, hospitals).

- Over 2,000 unique attributes were collected. These have been curated and harmonized into approximately 67 attribute categories.
- Of the 67 harmonized attribute categories, roughly 20 have substantial spatial and temporal coverage.
- The harmonization effort is still in progress, and further refinement will be needed.

2. DATA COLLECTION

2.1 SPATIAL AND TEMPORAL COVERAGE

The administrative units through which states reported information varied widely over time. State and county levels are the most common reporting units and show consistency across time, where available. In many states further geographic resolutions is available, such as zip codes, cities, and even individual facilities such as long-term-care facilities and hospitals. Overall, more than 13,000 unique locations appear in the database.

Figure 1 shows total data points collected over time by state, including all sub-state geographies, such as counties. A common pattern emerges of low data availability in March with an increasing amount of data becoming available starting in April. This corresponds to states reporting higher resolution information across the geographic, temporal, and demographic dimensions.

Also visible in many of figures throughout this report is a drop in total data points in the last day or two of collection. This is the result of lagged data reporting and how each state characterized the date of their data. Because of the hard cutoff in collection, if a state had not updated its reporting at the time collection stopped, the data for those dates did not get included. Although this slightly lowers the temporal coverage of the collected data, it also provides an interesting snapshot of how data reporting lags varied among states.

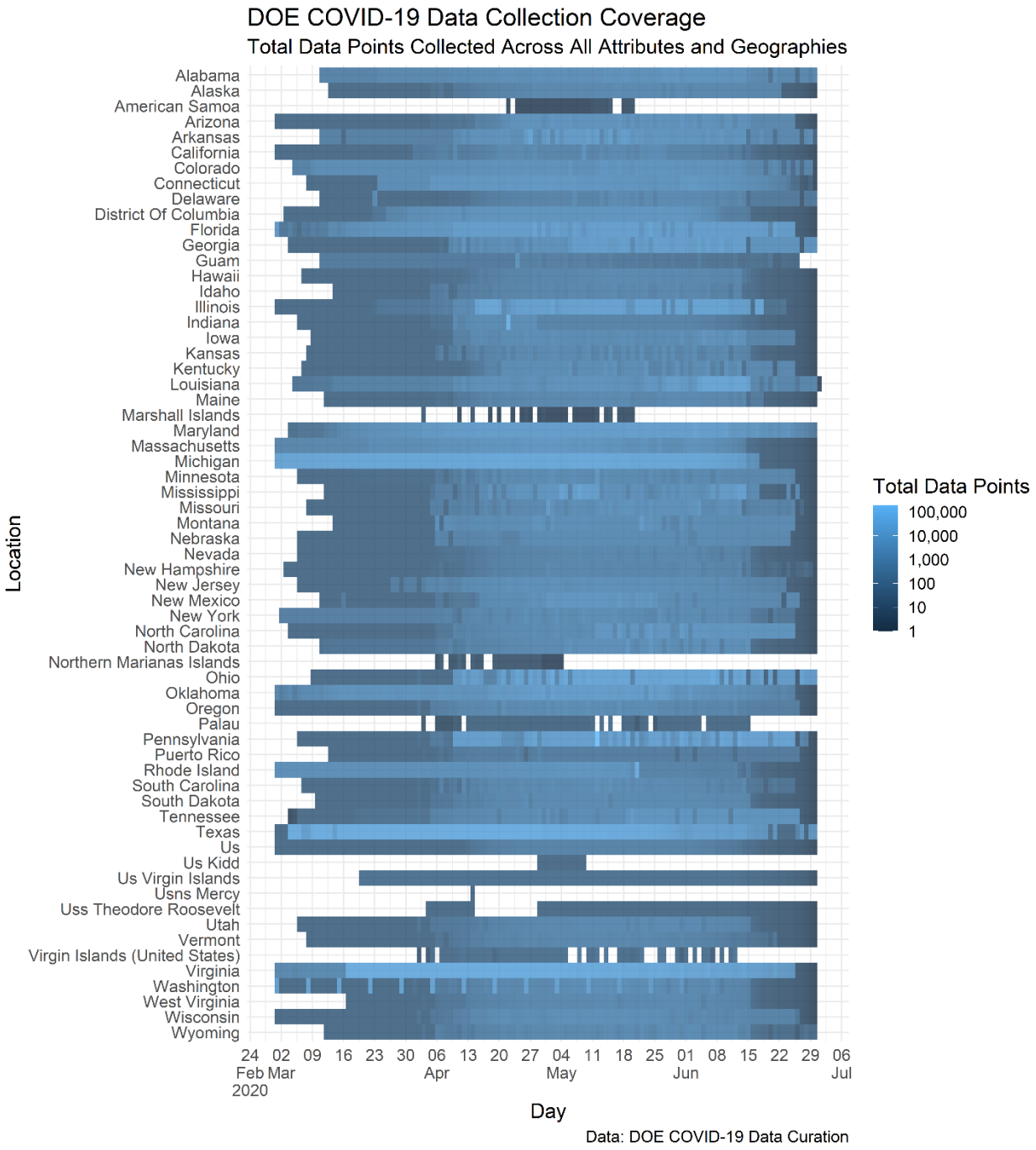

Figure 1. Total data points collected across US states and territories.

In general, states began reporting data at the county or county-equivalent level fairly quickly, as can be seen in Figure 2. Several states continued to report at this spatial resolution but also started to report data for other geographic units. Louisiana in mid-May started to subdivide counties into health regions while also reporting some attributes at an even higher spatial resolution such as census blocks. Another example is Pennsylvania reporting data for individual facilities, such as long-term-care facilities and hospitals.

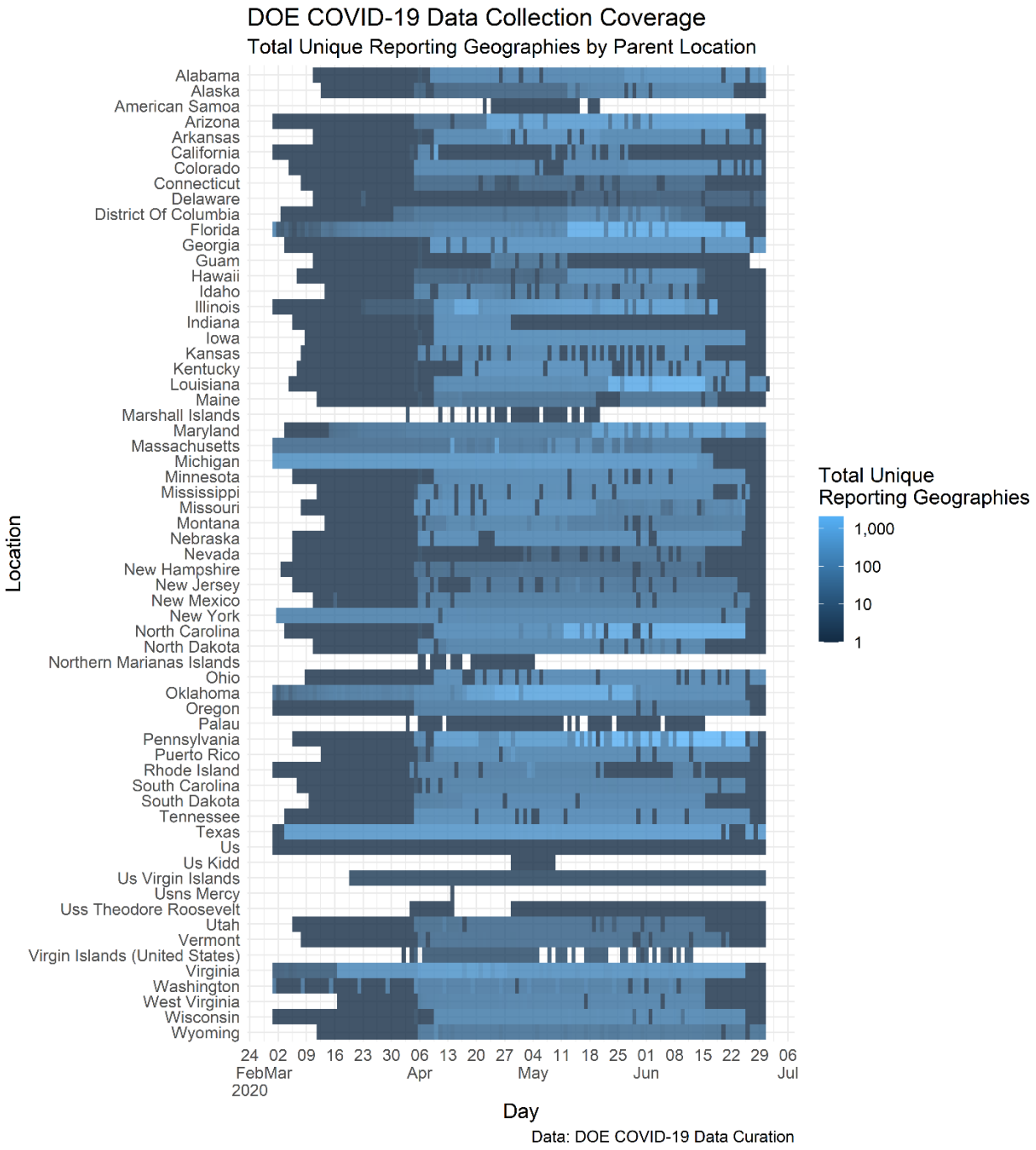

Figure 2. Total unique reporting geographies by US state and territory.

As might be expected, states varied considerably in how attributes were reported and categorized. An initial assessment yielded over 2,000 unique attributes in the database. From March into early June, only the broadest levels of information were being reported—total cases and total deaths were the only attributes consistently available. From June into the summer, more detailed information became available across each state, as can be seen in Figure 3.

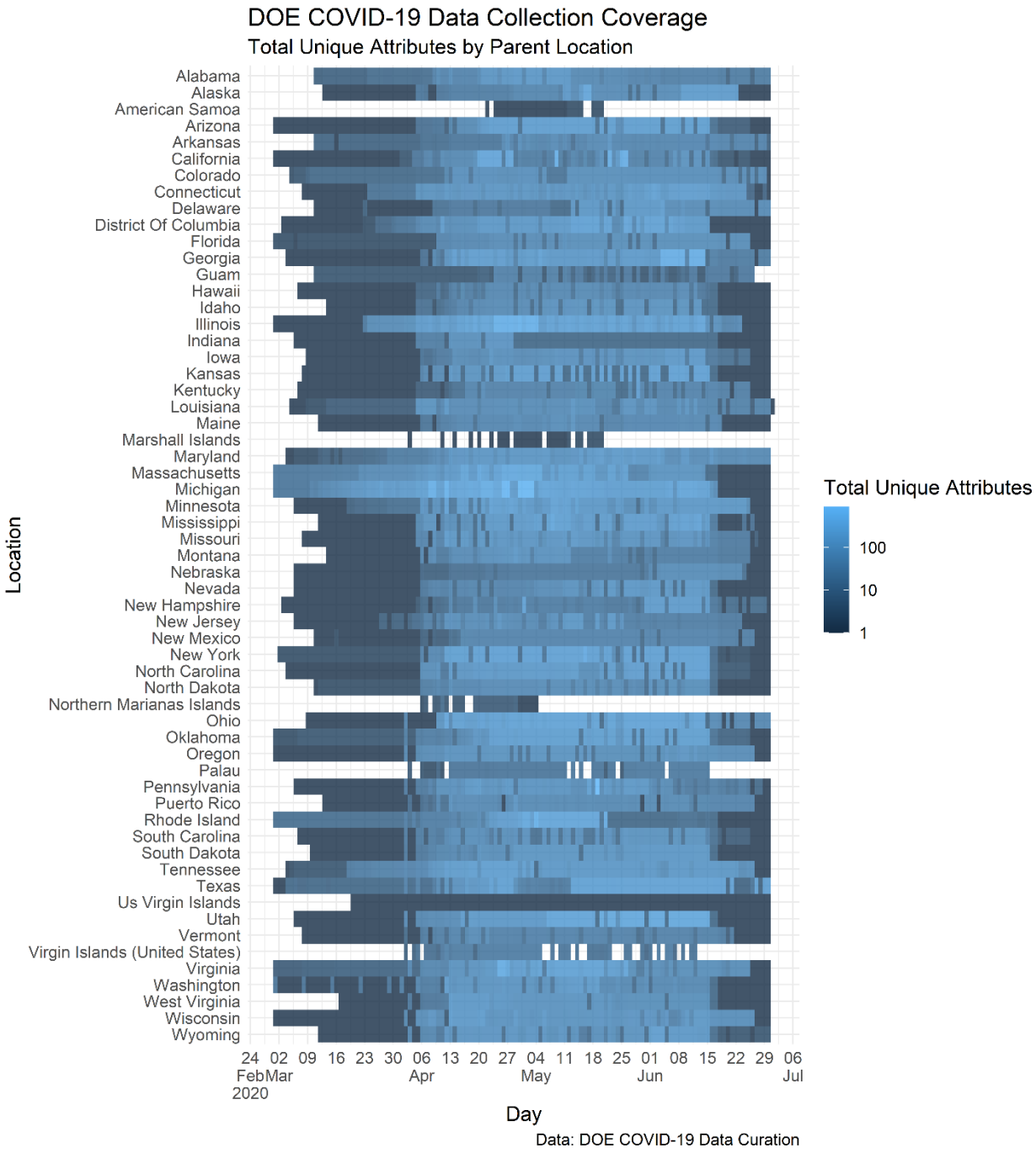

Figure 3. Total unique unharmonized attributes by US state and territory.

2.2 ATTRIBUTE COVERAGE

2.2.1 Attribute Harmonization

Harmonizing the raw reported attributes into consistent categories was a substantial effort that is still ongoing. The approximately 2,000 attributes have been categorized into 67 attribute categories that can contain multiple unique attributes.

For example, total cases (age 0–9), total cases (age 10–19), total cases (age 20–29), etc., are all individual attributes, but we have grouped them into a single attribute category of "Cases by Age Group". This structure, as well as harmonizing data in parent-child relationship such as long-term-care facilities, nursing homes, residential-care facilities, and correctional facilities collected under the parent attribute "facility", provides a much better picture of the attribute coverage available in the database.

Figure 4 shows which attributes categories have the most associated data across time and all geographies. Cases and deaths are some of the most common, as well as their breakdowns into age categories and other demographics.

Looking at the harmonized attribute categories across the most common geographic unit, US counties, narrows the available categories down to about 20 that have substantial spatial and temporal coverage. This can be seen in Figure 5.

The diverse state data sources required a harmonization effort to develop comparable attributes. This effort reduced more than 19,000 attributes to 1,900. The column on the left in Table 1 gives some of the largest harmonized parent attributes, with their report count in the column on the right. In the center column is a subset of child attributes associated with the parents; for example, any cases reported by age range are grouped under the Harmonized Parent "Cases by Age Range". Similar harmonization was completed on child attributes for ethnicity and race; these were combined as some states reported ethnicity and race separately, and others reported them as a combined attribute.

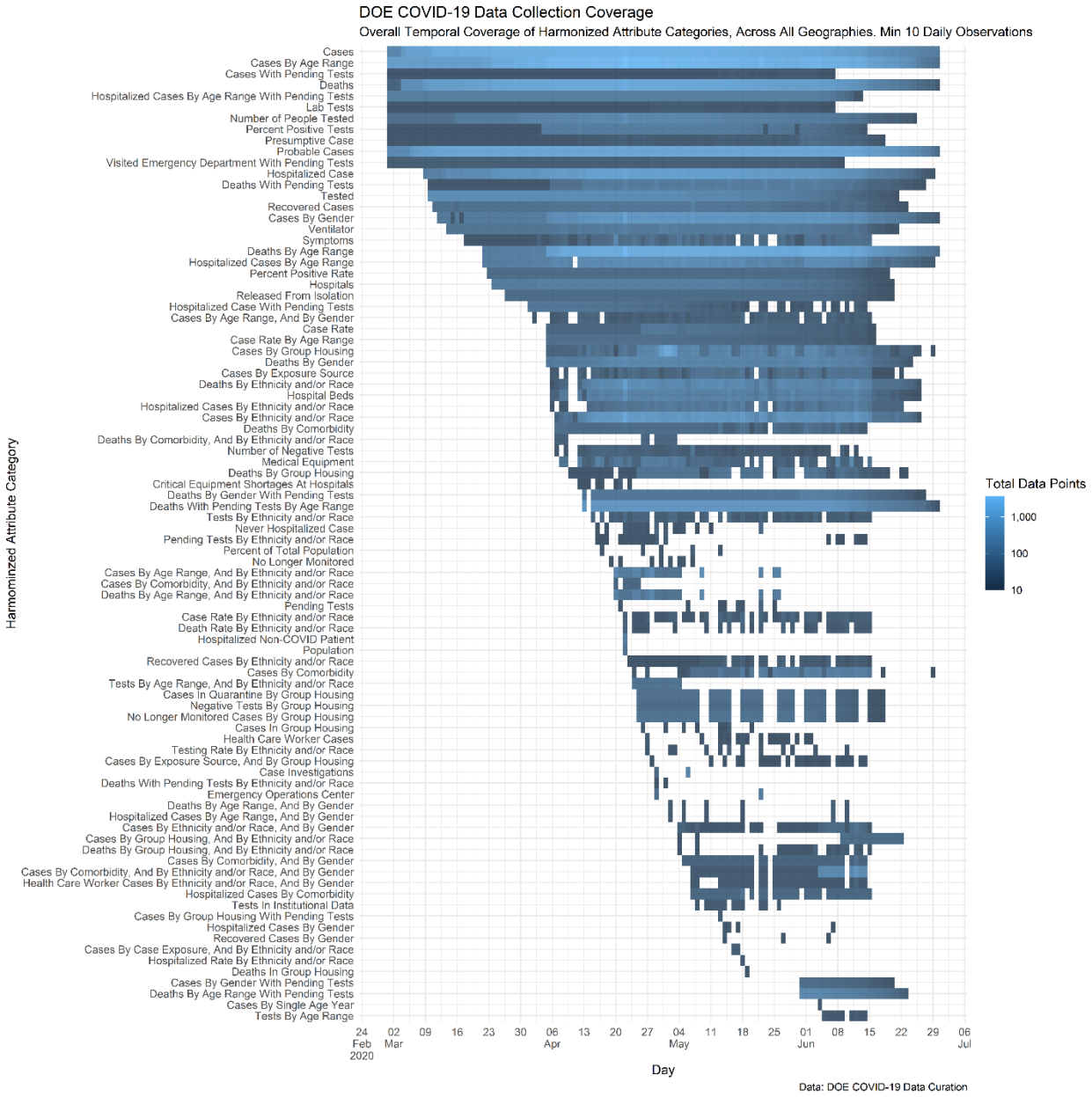

Figure 4. Most common harmonized attributes across all geographies.

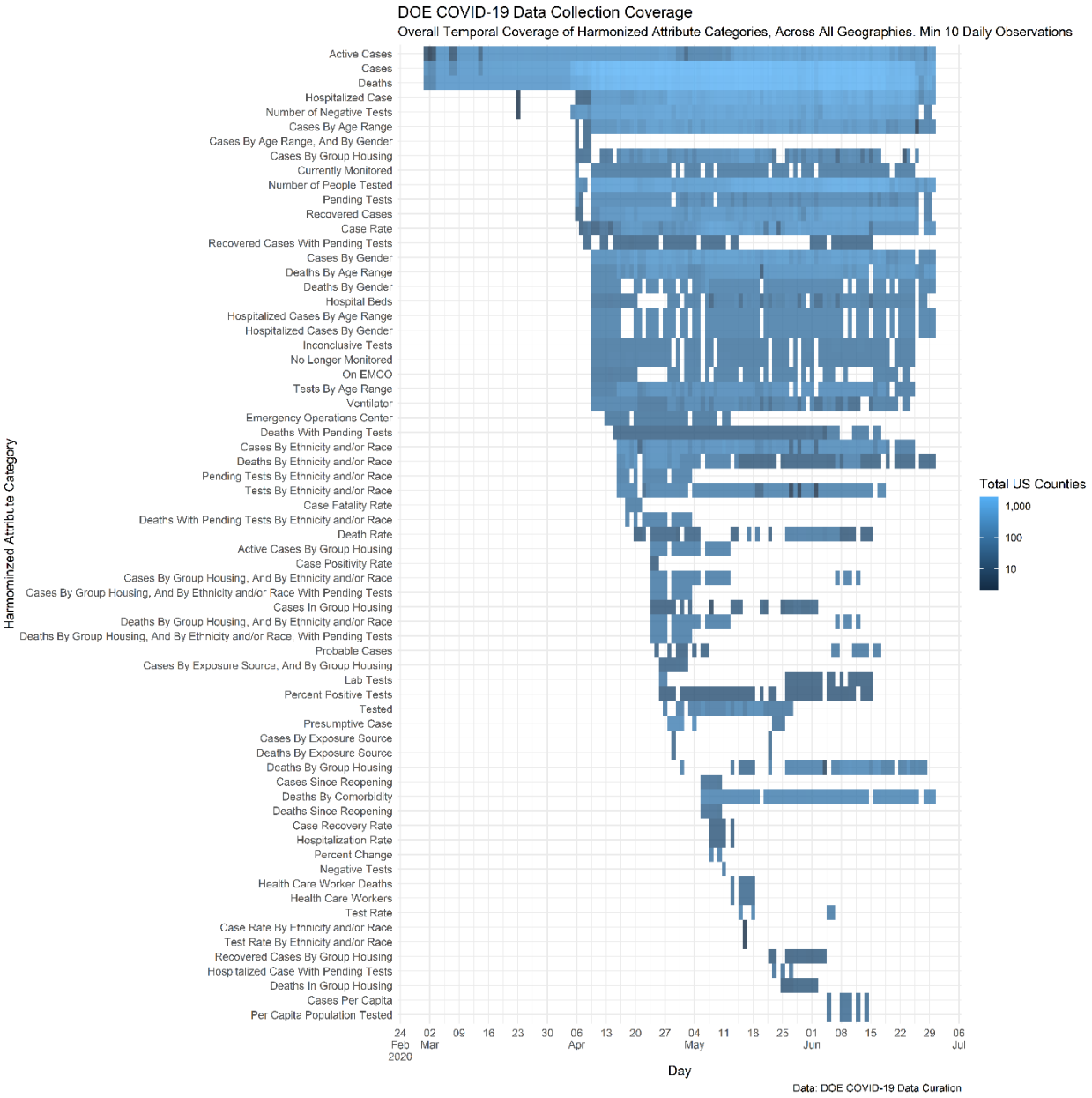

Figure 5. Harmonized attributes by total US county coverage.

Table 1. Most frequent harmonized attributes by count and example of child attributes.

Harmonized parent	Examples of child attributes	Count
Cases by Age Range	Age 50–64; Age 50–59; Age 45–64	225247
Cases	Tracking Data, Previous 24 Hours; Specific Date	153708
Probable Cases	N/A	106042
Deaths by Age Range	Age 0–9; Age >= 65; Age 20–29	91067
Deaths	Previously Hospitalized; Tracking Data, Previous 24 Hours; Presumptive Case	85768
Hospitalized Case	Adult; Currently Hospitalized; in ICU; on Ventilator	51350
Cases by Gender	Female; Male; Non-binary	30709
Cases By Ethnicity and/or Race	African American or Black; Hispanic/Latinx; Two or More Races	27945
Deaths with Pending Tests by Age Range	Age 0–9, Age 20–39, Unknown Age	26172
Tested	Commercial; PCR; Serology	19414
Hospitalized Cases by Age Range	In ICU; Pediatric Case, Median Age	18502
Hospitalized Cases by Age Range with Pending Tests	Age 5–17; Emergency Room Visit	18282
Deaths by Ethnicity and/or Race	Non-Hispanic/Latnix; American Indian or Alaskan Native	14490
Deaths by Gender	Female; Male; Non-binary	11856
Cases by Group Housing	Assisted Living; Behavioral Health; Correctional Facility	11240
Hospital Beds	Available Beds; Medical/Surgical Beds; ICU Beds	9462
Hospitals	Reporting; Using Surge Capacity	8624
Ventilator	Available; Alternative Ventilator, Pediatric Ventilator	8207
Deaths with Pending Tests	N/A	7395
Percent Positive Tests	Tracking Data, Previous 24 Hours, Tracking Data, Specific Day	7036
Recovered Cases	Tracking Data, Previous 24 Hours, Tracking Data, Specific Week	6389
Deaths by Gender with Pending Tests	Female; Male; Non-binary	5784
Hospitalized Cases by Ethnicity and/or Race	Asian; Native Hawaiian or Pacific Islander	5111
Cases by Comorbidity	Any Preexisting Condition; Diabetes; Pulmonary Condition	3596
Symptoms	Abdominal Pain; Fatigue; Nausea or Vomiting	3367
Medical Equipment	Coveralls; Distributed; Face Shields	3343
Released from Isolation	N/A	3308
Visited Emergency Department with Pending Tests	N/A	3098
Cases by Exposure Source	Community Transmission; Contact with Known Case; Travel	3089
Case Rate	Per 1000; Per 100,000	3008