DS 3001 Data Description

Kaitlyn Chou, Mackenzie Chen, Mrunal Kute

September 2025

1 Data Description

For this project, we plan to analyze the *COVID-19 Case Surveillance Public Use Data* from the Centers for Disease Control and Prevention (CDC) in order to create an algorithm to predict the trends of coronavirus disease 2019 (COVID-19) cases within the United States (US). Specifically, our goal or research question is to see whether the general trends of COVID-19 are increasing, decreasing, or stagnant in the future, as well as if the patient's underlying health issues affect this trend. This specific data set was created on May 15th, 2024 and was last updated on July 9th, 2024. While we had hoped to find more recent data, we were unable to find any data sets with data from the year of 2025. We suspect this is due to the fact that the year of 2025 is still ongoing, and therefore, datasets haven't been added yet for this year.

The key variables we are looking at in this dataset include:

- **Date Case** (`cdc_report_dt`) - When a specific case was reported to the CDC
- **Race/Ethnicity** (`race_ethnicity_combined`) - The racial background for each case
- **Underlying Health Condition** (`medcond_yn`) - Whether there is an underlying condition present for the case
- **Age Group** (`age_group`) - The age group of the specific case
- **Hospitalization Status** (`hosp_yn`) - Was the case hospitalized?
- **ICU Status** (`icu_yn`) - Was the case admitted to the ICU?
- **Death Status** (`death_yn`) - Did the case pass away?

Working with this dataset presents several challenges. Since we are using it for predictive purposes, it is important to determine how many cases resulted in death versus survival. As part of the analysis preparation process, we will

need to record the `death_yn` variable as a binary indicator, with 0 representing survival and 1 representing death. However, this dataset only captures mortality rates, meaning our ability to examine more nuanced outcomes, such as long-term respiratory complications is hindered. Specifically, underlying conditions might make a patient susceptible to death, but even if that patient got COVID-19 but didn't pass away, we still don't have a clear understanding of how COVID-19 affected them long-term. Thus, we will have to work around the lack of a variable for that when we do our analysis.

Our dataset, in general, is also quite big. It was quite difficult to export the file because it took a really long time. Additionally, we also have to account for the possibility that the dataset is not properly formatted, so we have to do some of our own cleaning and preparation to make sure the dataset is understandable. There is also a chance that there might be missing or an under reporting of data because some people might not report their case or follow up. Finally, some people could be biracial, so we have to account for that within that variable.

In order to clean the data, we will drop any null values, especially cases that have null values for hospitalization, death, or ICU status - chances are that these might be cases that were not followed up on. In addition, a lot of the "status" variables are in text format, which would have to change to an integer so that we can make it a binary (0,1) value. This will make it easier when reporting statistics on the data. Likewise, for more nuanced variables like "Race/Ethnicity", we will likely have to recode categories and standardize them, especially in cases where multiple ethnicities are included. For situations where "unknown" is marked for an ethnicity, that data will be isolated and analyzed separately. Further, we will check for "impossible data", any data that might have been mis-entered; for example, a negative age or a future date would indicate that we should drop the case. After cleaning the data, multiple sanity checks will be conducted and the data will also be plotted to ensure that the trends make sense.

This is the link to our dataset: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf/about_data.