IHME Assessment: Early COVID-19 Data

Research Scientist, Tobacco Metrics

Paul (Marty) Ross

May 31, 2023

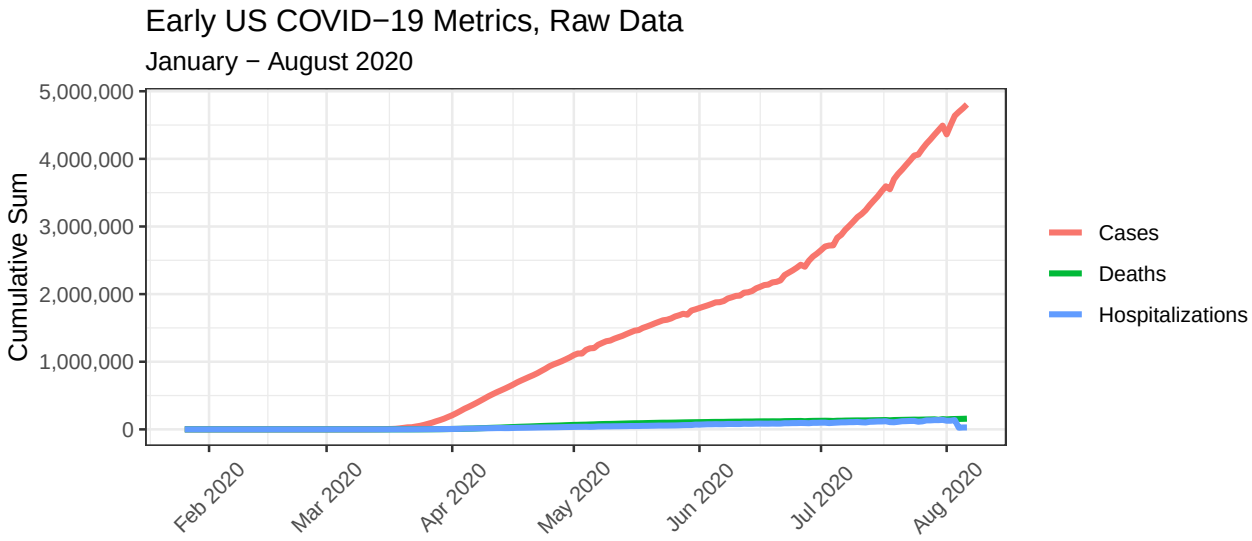
1 Cases, Hospitalizations, Deaths

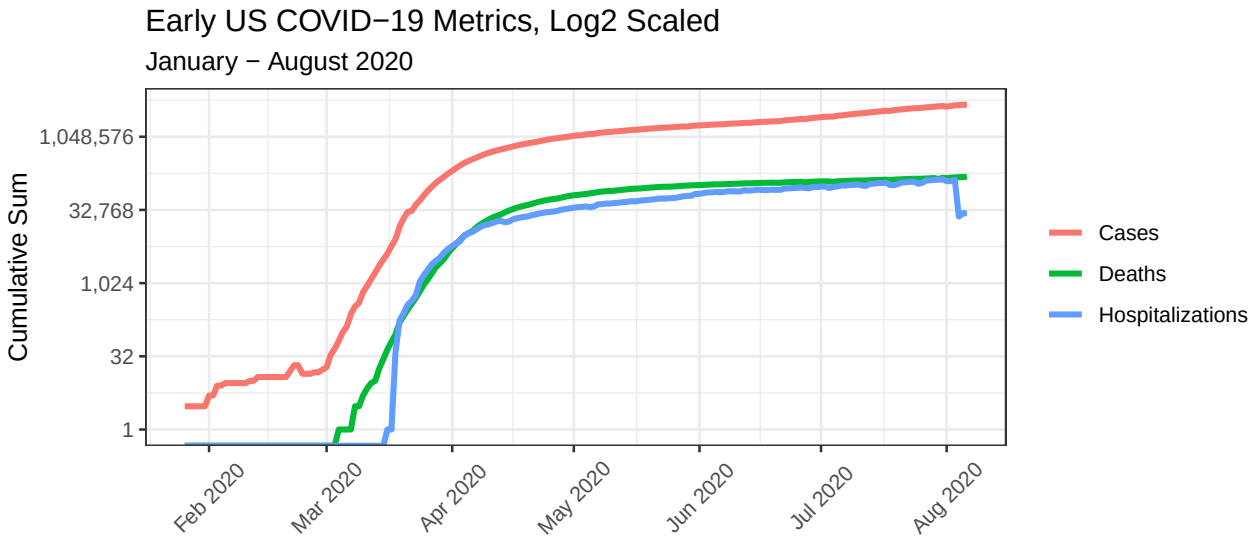
1.1 What is the relationship between cases, hospitalizations, and deaths? Describe how these indicators relate to each other, and visualize the relationships in at least 2 different ways.

During the COVID-19 pandemic, we saw cases, hospitalizations, and deaths rise in a temporally lagged relationship. In the *very* early stages of the pandemic, diagnostic tests for the novel SARS-CoV2 virus were in far too short supply to correctly enumerate the cases, and resulting hospitalizations and deaths. But when testing was widely available, the pattern observed during periods of surging cases was a roughly 2-week lag between upticks in cases and hospitalizations, then an additional 2-week lag before an uptick in deaths.

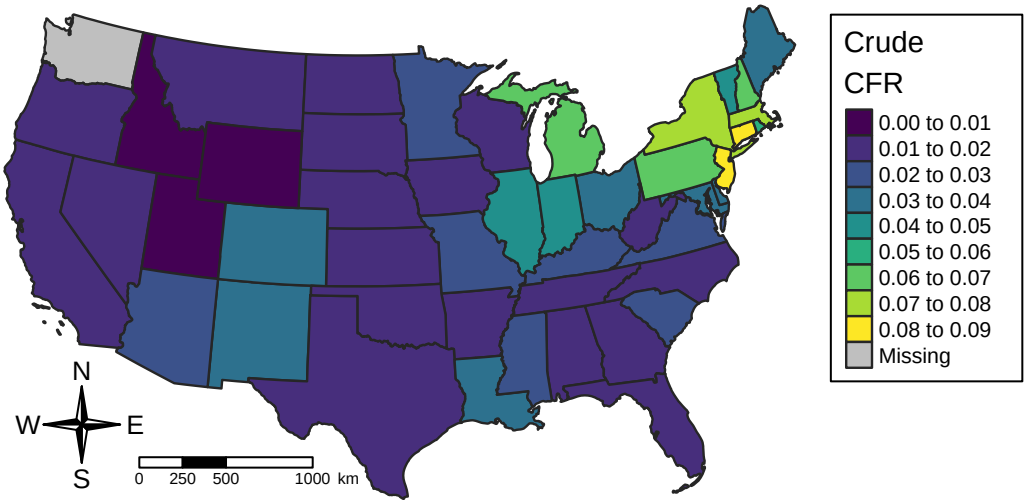
While not apparent in raw count space, in log space the difference between the cumulative death and cumulative case lines reflect the estimated 5% crude case fatality rate (CFR) of the original strain (~4-log difference).

Also interesting is the noisiness of the hospitalization data near the end of the plot, reflecting the reporting lag for the metric. As the hospitalization data is incomplete due to different legal obligations by state, we see lower levels of reported hospitalization than actually occurred. It lies near the level of deaths, where it should actually lie solidly between the cases and deaths.





A useful way to visualize the CFR across this dataset is by state in a chloropleth map. Of course this is a rough metric, and given the structural independence of health departments, divergent testing regimens, population densities, earlier geographic entry points, and differing levels of co-morbidities by state (smoking, obesity, etc.), we see a wide variety of CFR's, which in aggregate yield a ~5% crude CFR nationally. This figure is inflated due to undertesting, but as a crude measure it is useful for exploratory analysis.

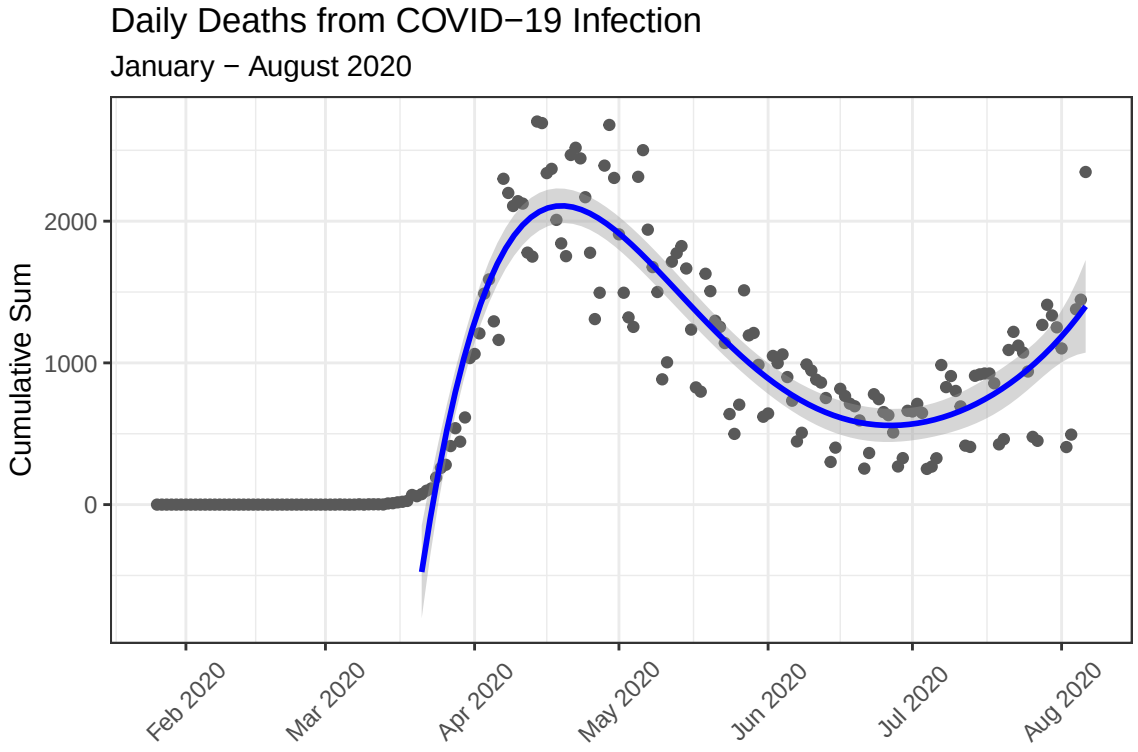


2 Fit Daily Death Metric

2.1 Fit a curve of daily deaths, utilizing these inputs. Describe the approach you used and visualize the results.

This was a challenge to ensure the daily death totals were accurate, requiring imputing data for missing days with the cumulative total of the previous day. I fit the daily deaths using a 5^{th} degree polynomial in a linear model using `lm`, allowing adequate degrees of freedom to get a tight fit to the data and capture the local mins and maxes of the first wave and beginning of second wave of the pandemic. Additionally, I constrained the model to begin on March 21^{st} to ignore the low-count space and poor testing capacity of the early pandemic.

The model performs well, falling between the cycling high and low values, which reflect the weekend reporting lag that happened in many jurisdictions.



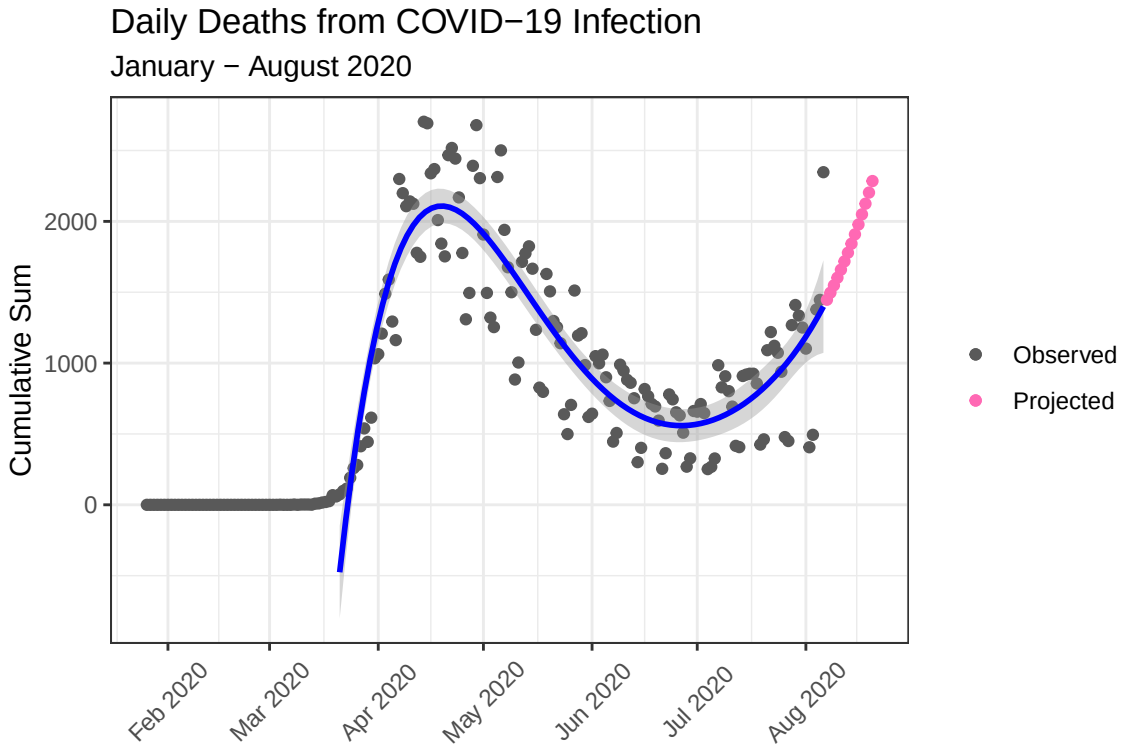
3 14-day Model Projection

3.1 Create projections for 14-days after the last observed data point. Visualize the result. Describe the benefits and limitations of your approach. Where do you think this approach has performed particularly well? What types of situations cause your model to struggle?

Overall, I trust the projections of this model to perform well over the course of the following 14-days, as the dataset ends at a period of rapid, near-logarithmic increase. We would expect the daily deaths to increase at something like this rate.

Some limitations of the 5^{th} degree polynomial are that it is likely over-fitting data that we know to be incomplete. Even bounding the model from March 21^{st} includes a large amount of early data where broadly available testing was still coming online. The literature on the subject points to a significant undercount of cases and deaths due to COVID-19 at this stage.

Additionally, using the polynomial fit would do a poor job of fitting at other stages of COVID-19 case surges, such as predicting a peak and downturn of daily deaths.



4 Follow-up

4.1 Lastly, describe future areas of exploration or improvement for your approach. If you had more time, what would you do next?

As I really enjoy incorporating geospatial approaches to public health problems, I would love to do a few things.

1. Build a spatial regression, incorporating demographic (race, SES), behavioral (tobacco use) and environmental (air quality) data to identify significant risk factors and covariates that yield the differences observed in crude CFR state-by-state.
2. As I have population data, I would like to create a companion chloropleth map of relative risk by state, using the cumulative case total to estimate expected deaths by population. This would very simply visualize where we see higher rates of death normalized per capita.
3. I'd love to build a quick Shiny app, as I think such approaches are powerful narrative tools for education, insight, and discovery. This is similar to my recently completed wildfire-oriented Master's capstone.

{Click for Wildfire Smoke App}

Thanks for this opportunity! - Paul (Marty)