Analysis plans

Our overall research question is whether past reported COVID-19 cases might be a useful input to forecasting hospitalizations.

There are two common methods for recording case data: (1) by the date of a laboratory test confirming that an individual has COVID-19, which we refer to as the *test date*; and (2) by the date that a public health agency released data including that individual's positive test result, which we refer to as the *report date*. Of these, the test date is the more epidemiologically relevant signal; the report date is always at least one day after the test date, and may be substantially later. These delays in reporting do not reflect any meaningful information about the disease transmission process. However, there are multiple challenges with working with test date date. First, most COVID-19 data aggregation systems collect data by report date. Second, test date data is subject to a substantial backfill process, in which data for past dates are updated as new records are released. There are some similar adjustments to report date data, but they are much less systematic; typically, once the data report for a particular date is made, it is not subsequently updated.

Models

In an investigation of the utility of case data for forecasting hospitalizations, it is necessary to carefully consider whether we use cases by test date or by report date. If we are using test date data, we must also consider the impact of the reporting backfill process on forecast skill. To address these questions, we formulate five model variations. All of these use past hospitalizations as an input to forecasting; four of them use different variations on case data and different strategies for addressing partial reporting for those case data.

So that we can focus on understanding the importance of the different variations of case data on modeling hospitalizations, all of the models below will use hospitalization data retrieved at a single point when the analysis was run. While this ignores possible small revisions that occured since the data were initially reported, this focuses the analysis on the question at hand. (We could provide some summary stats about revisions of hosp data in MA.)

1. **Solo**: A reference model that forecasts hospitalizations using only the past reported values of hospitalizations as an input.
2. **ReportCaseRealTime**: In addition to past hospitalizations, this model uses report date cases as an input to forecasting hospitalizations.
3. We consider three variations on models using test date cases as an input to forecasting hospitalizations, with different strategies for handling the backfill process:
    a. **TestCaseFinal**: This model uses "final" values of test date cases as an input. In practice, we use the reported test date cases from (some time when values have stabilized).
    b. **TestCaseRealTime**: This model uses the values of test date cases that would have been available for forecasting in real time, with no adjustments for partial reporting.
    c. **TestCaseNowcast**: This model uses nowcasts of test date cases as an input to forecasting hospitalizations. These nowcasts combine the real-time reports of test date cases as an input to a model that predicts the final reported test date cases.

We formulate more specific research questions about the utility of COVID-19 cases as an input to forecasting hospitalizations in terms of comparisons of the forecast skill of these models:

- Are report date cases helpful for forecasting hosps? Compare models 1 and 2
- In an ideal setting without reporting delay, is the more-epidemiologically-relevant signal of test date cases helpful for forecasting hosps? Compare models 1 and 3a, and/or models 2 and 3a.

- In the real-world setting with reporting delay, is directly using as-of reported cases helpful for forecasting hosps? Compare models 1 and 3b, and/or 3a and 3b
- In the real-world setting with reporting delay, is using as-of reported cases adjusted for that reporting delay helpful for forecasting hosps? Compare models 1 and 3c, and/or 3a and 3c, and/or 3b and 3c

Other thoughts

- We probably shouldn't put any time into thinking about actual nowcasting implementations until we've done the 1-3a and/or 1-3b comparisons and seen something indicating this could be useful.
- If a message like "using as of test date cases isn't super helpful, but using final reported test date cases would be helpful if they were available; nowcasting is a possible future direction to close this gap" is good enough, we could leave it out of this work

Forecast targets

In this analysis, we focus on having models make forecasts of several specific targets with public health relevance.

- Predictions of 1- to 30-day ahead incident hospitalizations. These forecasts will be structured identically to the COVID-19 Forecast Hub, to facilitate comparisons with those models if desired. These forecasts will be generated retrospectively, once for ever week, using data that were available as of each Monday of a given week.
- Predictions of the date at which incident hospitalizations reach their peak, for selected waves. Specifically, we will assume that forecasts for peak hospitalizations were solicited for the late-summer Delta surge in Massachusetts from mid-July 2021 through mid-October and then again for the Omicron surge from mid-November 2021 through early February 2022.
- [should we include peak value as well?]

Ground truth data for incident hospitalizations are taken from the HealthData.gov (HHS) source. [It probably makes things much simpler to use these data as ground truth for everything, but I am tempted to consider a smoothed version, say centered moving window of current day +/- 3 days on either side, of the daily hospitalization data as the ground truth target for peak timing.] The peak day will be defined as the day on which the peak number of new incident cases was observed. In case of a tie, the day with a higher centered moving average (+/- 3 days on either side) will be chosen as the day of the peak.