# 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.

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.

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" reported 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 reported 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