Guidance for the CDC COVID-19 Hospitalization Forecasting Project
First released: 2 April 2020

Table of Contents

Objective

The US CDC will host the COVID-19 Hospitalizations Forecasting Project, open to all, with the goal of generating probabilistic forecasts of COVID-19 activity in the US during 2020. For each week between April 6th, 2020 through August 29, 2020, participants will be asked to provide national-, state-, or county-level probabilistic forecasts for a set of specific targets related to the trajectory of COVID-19 related hospitalizations. Forecasts that are generated conditional on specific assumptions regarding future mitigations, interventions, and policies will be formally identified as such through model descriptions and metadata. Data from both individual and ensemble forecasts (weighted combinations of multiple submitted models) will be shared with decision-makers at CDC.

Eligibility

All are welcome to participate in this collaborative challenge, including individuals or teams that have not participated in previous CDC forecasting challenges. Teams do not need to provide forecasts for all locations or targets to participate in the project. Teams may provide point forecasts or probabilistic forecasts, and will be encouraged to provide both for whichever targets and locations they choose to model.

Overview

As of the time of writing this document, reliable long-term data streams for data related to COVID-19 hospitalizations are still emerging. Some sites (e.g. COVID Tracker) are tracking and aggregating data on COVID-19 hospitalizations and tests from state departments of health. Additionally, the CDC is activating a COVID-19 surveillance network of hospitals, similar to FluSurv-Net. Data from this new source, COVID-Net, will be available back to March 1, 2020, and will have age-specific hospitalization rates estimated for the national level from a small subset of representative hospitals.

Data

It is anticipated that one or more sources will emerge as reliable reference points for hospitalization data by state. Forecasts for this project will focus on predicting new incident hospital admissions. However current data sources and reports do not always distinguish between new incident hospitalizations, prevalent hospitalizations, and hospitalizations reported

in a given week. As new data sources become available and existing data sources become more standardized, the organizing team will work to provide additional guidance about which data sources are considered to be the gold standard. At the time of writing, the COVID Tracker website is providing a clear measure of “cumulative” incident hospitalizations by state.

Interventions
Teams whose models can factor differing levels of intervention effectiveness into their forecasts are welcome and encouraged to submit forecasts from multiple different models that make different assumptions about future intervention strategies.

To keep track of such different assumptions, we pose the following set of questions for each model, answers to which are tracked in the metadata file documented below.

1. Is this forecast conditional on specific assumptions regarding future mitigations, interventions, or policies (including the assumption that mitigations, interventions, or policies remain unchanged)?
2. If a model indicates "Yes" to the above, under what specific such assumptions are these forecasts valid?

Teams are encouraged to submit multiple forecasts under different scenarios if their modeling framework can accommodate this.

Locations
As many decisions are made on local scales, state-level forecasts may hold the most operational value. Forecasts will be accepted for the national level (FIPS code = “US”) for the state level (FIPS code 2-digit character string), and county level (FIPS code 5 digit character string, with first two characters representing the state FIPS code). A file with FIPS codes for states and counties is available through the `tigris` R package, and saved as a public CSV file.

As requested by specific teams or decision-makers, the organizers may create groupings of county-level FIPS codes (e.g. into a specific metro area) for which forecasts can be provided.

Targets
Throughout this description, and in the templates, we use the standard definition of “epidemic weeks” (EW) or “MMWR weeks”, as defined by the CDC and other public health agencies. There are standard software packages to convert from dates to epidemic weeks and vice versa. E.g. MMWRweek for R and pymmwr and epiweeks for python.

This template defines a set of targets relative to the time series of observed new COVID-19 hospitalizations by state and county in the US. The targets are described in greater detail below in the Target Details section. In brief, the targets are:

- 1- through 25-week ahead incident COVID-19 new hospitalizations

    Teams are encouraged to submit forecasts for the set of horizons for which they feel confident in their model forecasts. There may be reasons for choosing a

specific horizon for a specific model. However, in the absence of a clear specific choice, we encourage teams to submit either 1-2 weeks, 1-6 weeks or 1-25 weeks to assist with standardization.

- "Peak hospitalizations"

  The highest number of new COVID-19 hospitalizations observed within the range of 2020-ew10 and 2020-ew35. The peak is defined as the highest number observed within these weeks for a given location. If a distinct peak occurs early in the specified weeks, project organizers may "reset" the time-frame to allow for teams to forecast a second peak.

- "Peak week"

  The week of the first new COVID-19 hospitalization peak between 2020-ew10 and 2020-ew35. The peak is defined as the highest value observed within the given time-range. If an early distinct peak occurs, project organizers may "reset" the time to allow for teams to forecast a second peak.

Forecasts should be submitted on Thursdays by 11:59pm ET. Teams may start submitting forecasts at any time during the project time-period, although early participation is encouraged. We have provided a table showing the dates on which forecasts will be due (typically Thursdays), and, for each forecast date, which weeks the "week-ahead" targets should correspond to. (**For teams participating in the ILI forecasting project, please note that the definition of the week-ahead forecasts is different in this project.)** For complete instructions on submitting forecasts, please see the Template and formatting details section below.

Ensemble forecasts will be created by synthesizing submitted forecasts into a single aggregated distribution. Project organizers, in collaboration with CDC colleagues, will determine appropriate weighting algorithms.

Target details

1- through 25-week-ahead new hospitalization count

- Type of target: discrete
- Description: The number of **new** hospitalizations for {1, 2, 3, …, 25} week(s) after the due date. Please see the dates table for exact information on which EW week each target should refer to.
- Units: cases, i.e. non-negative integers
- Bin boundaries: by 50 from 0 to 500, 100 from 500 to 10,000, and by 1000 from 10,000 to 100,000, i.e. {0, 50, 100, 150, …, 450, 500, 600, 700, …, 9900, 10000, 11000, … 99000, 100000}. The last bin is assumed to encompass 100,000 and everything higher.

Peak week

- Type of target: date
- Description: This target captures information about the epidemic week (defined using MMWR week standards) in which the reported new COVID-19 hospitalizations for a

given location will achieve its highest value between the 2020-ew10 (start date March 1, 2020) and 2020-ew35 (start date August 23, 2020).
- Units: week
- Categories/bins: Point predictions and categories for probabilistic distributions will be represented by an unambiguous notation for epidemic weeks (e.g., “2020-ew33”). The set of valid values for this target are therefore {“2020-ew10”, “2020-ew13”, …, “2020-ew34”, “2020-ew35”}.

Peak hospitalizations
- Type of target: discrete
- Description: This target contains information about peak value of new hospitalizations observed in a given location between the 2020-ew10 (start date March 1, 2020) and 2020-ew35 (start date August 23, 2020).
- Units: cases, i.e. non-negative integers
- Bin boundaries: by 50 from 0 to 500, 100 from 500 to 10,000, and by 1000 from 10,000 to 100,000, i.e. {0, 50, 100, 150, …, 450, 500, 600, 700, …, 9900, 10000, 11000, … 99000, 100000}. The last bin is assumed to encompass 100,000 and everything higher.

Template and data formatting details

Teams and models

Teams interested in participating in the CDC COVID-19 Hospitalization Forecasting Project can submit forecasts from multiple models. Teams are encouraged to provide forecasts for locations and targets for which they feel their models are well-suited. Teams with multiple models that make distinctly different assumptions about interventions are encouraged to submit these separately. Teams with a group of models that make similar assumptions about interventions can choose to either submit them as a single ensemble model or as separate model entries.

Prior to the first submission for a given model, the submitting team must provide a metadata file with structured information about the model. Each submitting team must choose a full name and an abbreviation for both their team and their model to uniquely identify their submissions.

The metadata file for each model must be named `metadata-[teamabbr]-[modelabbr].txt` and include the following information:
- team name
- team abbreviation for submission files (<20 characters, alpha-numeric and underscores only, no spaces or hyphens)
- model name
- model abbreviation (<20 characters, alpha-numeric and underscores only, no spaces or hyphens)
- model contributors, main point of contact(s) should have an email specified
- brief description of each data source
- whether or not the model itself is a type of ensemble model

- a binary Yes/No indicator of whether this model is conditional on specific assumptions regarding future interventions.
- a brief description under what specific such assumptions are these forecasts valid.
- methodological description, including citations if appropriate.

An example metadata file is provided.

Forecast file format
In what follows, we refer to a "forecast" as a collection of quantitative predictions that are specific to a location and target specified above. One forecast can be submitted for a given model on or before the forecast "due date" specified in the dates table. A forecast consists of a single plain-text file, in a particular format, that encapsulates the set of predictions for all or a subset of locations and targets.

A forecast may be submitted using the template for hospitalization forecasts, details of which are provided below. (Please note the linked template has predictions for 3 locations. In general a forecast file can provide forecasts for as many locations and targets as is desired.) In general, a prediction for a specific location and target will be specified by point forecasts and a binned representation of a probability distribution. We will refer to these two representations as "point forecasts" and "bin forecasts". Teams may submit only one or the other type, although they are strongly encouraged to submit both.

Forecasts are encouraged to provide probabilistic forecasts (i.e., probability 0.5 peak will occur on week 2; probability 0.3 on week 3, etc...) as well as point predictions for each target. The probabilities for each single probabilistic prediction should be non-negative and sum to 1. If the sum is greater than 0.9 and less than 1.1, the probabilities will be normalized to sum to 1.0. If any probability is negative or the sum is outside of the 0.9-1.1 range, the forecast will be discarded.

Here is a data dictionary describing the columns in the forecast template:
- location: location code for the prediction
- target: target for the prediction
- type: the type of prediction, should be either "point" or "bin"
- bin: the lower bound of the "bin" of the empirical distribution
- value: the numeric point prediction or bin probability

Since character values can appear in every column (e.g. `2020-ew25` is valid data for the bin or value column, FIPS values may have leading zeros) please take care to retain the data in these columns in character format.

All forecasts should be structured to match the hospitalization template. The column structure of the template should not be modified in any way. Rows for targets or locations that have not been forecasted should be left out. Rows for bins with zero probability may also be left out to save space. Peak height and week-ahead forecasts should be given in the provided intervals

labeled “bin” on the submission sheet. For example, the row with bin==100 represents the probability that the target will eventually be observed to be in the interval [100, 150). The probability assigned to the final bin labeled 100000 includes the probability of new hospitalizations being greater than or equal to 100,000, or in the interval [100000, \infty].

Forecast file name
A forecast submission using ILINet data through epiweek 12 submitted by John Doe University (team abbreviation: JDU) for the Deep Learning Special Sauce model (model abbreviation: DLSpecialSauce) on Thursday, April 9th, 2020, should be named “2020-ew15-JDU-DLSpecialSauce.csv” where 2020-ew15 is the corresponding value in the `forecasts_due_ew` column from the dates table. The 1-week ahead forecasts contained in this file would refer to 2020-ew16, as indicated by the `forecasts_1_wk_ahead` column in the dates table.

Forecast file storage and submission
Submitted forecasts will be stored in a private GitHub repository. A public version of the repository will contain the updated guidelines and ensemble forecasts that combine multiple models into a single model. The public repository lives at:
https://github.com/cdcepi/COVID-19-ILI-forecasting
To gain access to the private repository, teams who intend to participate in the challenge should send their model metadata file to the project organizers at flucontest@cdc.gov along with the GitHub username of the participant who will be submitting forecasts. We request that all forecast submissions for each team be submitted via a GitHub pull request. We will provide instructions for submitting via pull request if this process is new for a team. As a backup, teams that are unable to submit via a pull request may email their submission files to: flucontest@cdc.gov.

In the COVID-19-ILI-forecasting private repository, there are three main folders for storing forecasts:

- state-forecast-data
- nation-region-forecast-data
- hospitalization-forecast-data

Each of these folders will contain subfolders for each model for which forecasts are being submitted. The subfolders will follow the naming convention of `[teamabbr]-[modelabbr]`. Subfolders will contain the metadata file for that model and all submitted forecasts for that model.

For example, for the JDU team and DLSpecialSauce model for state forecasts, the metadata file would have the path:
`hospitalization-forecast-data/JDU-DLSpecialSauce/metadata-JDU-DLSpecialSauce.txt`
And the forecast submitted during 2020-ew15 would have the path:
`hospitalization-forecast-data/JDU-DLSpecialSauce/2020-ew15-JDU-DLSpecialSauce.csv`

Forecast licensing

At an appropriate time, the complete data repository will be made available online and archived in a permanent and public data repository under a Creative Commons 4.0 license license, with a DOI, to facilitate future use and citation/referencing. Ensemble forecasts, that aggregate all forecasts received, may be made public sooner than team-specific forecasts. Teams will have the opportunity to make their forecasts anonymous before they are made public. A collaborative academic manuscript describing this forecasting project will be coordinated by a designated representative of CDC.

Data Sources

Historical national surveillance data may be used for training and model development. Historical FluSurv-Net data (this is different from COVID-Net data) is available at https://gis.cdc.gov/GRASP/Fluview/FluHospRates.html. Additional data resources will be added here as they become available. These data are typically updated every Friday at noon Eastern Time. The cdcfluview package for R can be used to retrieve these data automatically.

Teams are welcome to utilize additional data beyond CDC data - additional potential data sources include but are not limited to: Carnegie Mellon University's Delphi Group's Epidata API, the COVID Tracker website, and Health Tweets. The Epidata API includes weekly surveillance data as they were first published and in their most up-to-date version following revisions to initially published data.

Contact Info

Additional questions may be addressed to flucontest@cdc.gov.
