The COVID-19 Forecast Hub and covidHubUtils: curating and evaluating COVID-19 forecasts for hospitalizations, cases, and deaths in the United States

Estee Y Cramer[1*], Yuxin Huang[1*], **Yijin Wang**[1*], Evan L. Ray[1], Matthew Cornell[1], Johannes Bracher[2,3], Andrea Brennen[4], Alvaro J Castero Rivadeneira[1], Aaron Gerding[1], Katie House[1], Dasuni Jayawardena[1], Abdul Hannan Kanji[1], Ayush Khandelwal[1], Khoa Le[1], Jarad Niemi[5], Ariane Stark[1], Apurv Shah[1], Nutcha Wattanchit[1], Martha W Zorn[1], Nicholas G Reich[1], on behalf of the US COVID-19 Forecast Hub Consortium

[*] Equal contribution [1] University of Massachusetts Amherst, [2] Karlsruhe Institute of Technology and Heidelberg Institute for Theoretical Studies, [3] Computational Statistics Group, Heidelberg Institute for Theoretical Studies, [4] IQT Labs, [5] Iowa State University

Introduction

To understand how the COVID-19 pandemic would progress in the United States, dozens of academic research groups, government agencies, industry groups, and individuals have been producing forecasts for COVID-19 outcomes.The Forecast Hub was launched in early April in 2020. In coordination with the US CDC, the Forecast Hub curates an **authoritative, real-time record** of forecasts about reported COVID-19 **cases**, **hospitalizations**, and **deaths**. As of November 15th 2021, we have collected nearly **76 million** individual point or quantile predictions contained within over **5,238** submitted forecast files from over **108** unique models.

We developed a unique set of workflow and software infrastructure to curate, store and analyze all forecast data in a standardized fashion. The open-sourced dataset and softwares are widely accessible to the public, thus fostering a transparent, open science approach to support public health efforts.

Goals

The COVID-19 Forecast Hub aims to:

- Provide decision-makers and general public with reliable information about where the pandemic is headed in the next month;
- Assess reliability of forecasts and gain insight into which modeling approaches do well;
- Create a community of infectious disease modelers underpinned by an open-science ethos.

Data Collection

Each week, the COVID-19 Forecast Hub receives forecasts of deaths, cases and hospitalizations in the US due to COVID-19. These forecasts are submitted for varying locations and horizons:

Outcome	Scale	Locations			Horizons Stored	Number of quantiles for probabilistic forecasts	Earliest Forecast Date
		County	State	National			
Incident Cases	Weekly	X	X	X	1 - 8 weeks	7	2020-07-05
Incident Hospitalizations	Daily		X	X	1 - 130 days	23	2020-03-27
Incident Deaths	Weekly		X	X	1-20 weeks	23	2020-03-15
Cumulative Deaths	Weekly		X	X	1-20 weeks	23	2020-03-15

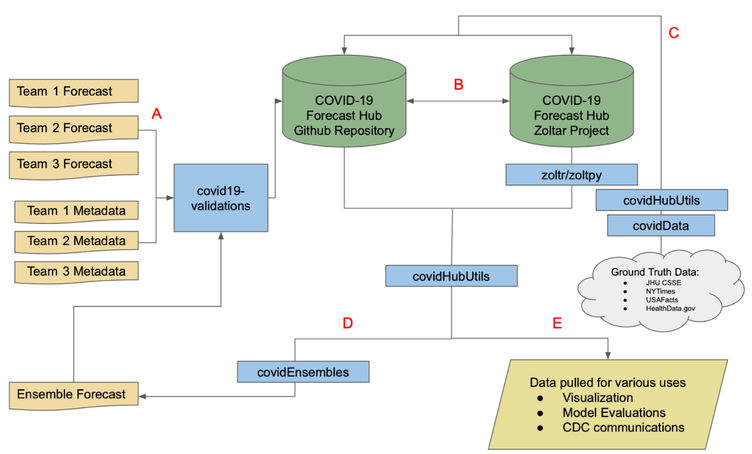
The Forecast Hub served as a public, independent registration system for these forecast model outputs.

Approach

We created an ecosystem of infrastructure to establish a standardized and comparable set of forecast data from modeling teams. The schematic of this workflow is the following:

(A) Forecasts are submitted to the COVID-19 Forecast Hub GitHub repository and undergo data format validation.
(B) A continuous integration service ensures that the GitHub repository and Zoltar database stay in sync
(C) Truth data for visualization, evaluation, and ensemble building are retrieved once per week using both the *covidHubUtils* and the *covidData* R packages.
(D) Once per week, an ensemble forecast submission is made using the *covidEnsembles* R package.
(E) Using the *covidHubUtils* R package, forecast and truth data may be extracted from either the GitHub or Zoltar database in a standard format for tasks such as scoring or plotting.

Workflow

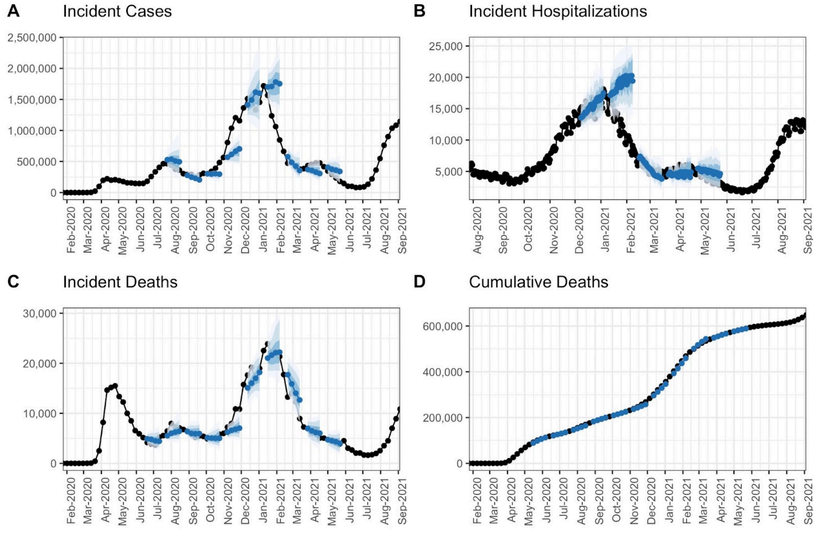


covidHubUtils: Query & Visualization

We have collaborated on *covidHubUtils* R package to facilitate data analysis using forecast data collected in the Forecast Hub. It could:

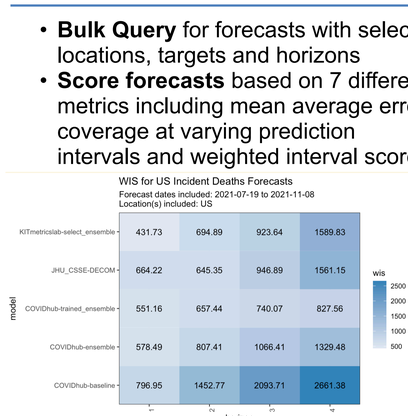
- **Query forecast** data from Zoltar or COVID-19 Forecast Hub GitHub repository with specified parameters of locations, forecast dates, targets, horizons, as of dates;
- Download and process raw version of **ground truth data** from sources such as JHU CSSE, NYTimes, USAFacts and HealthData.gov;
- **Plot** preloaded forecast data with truth data.

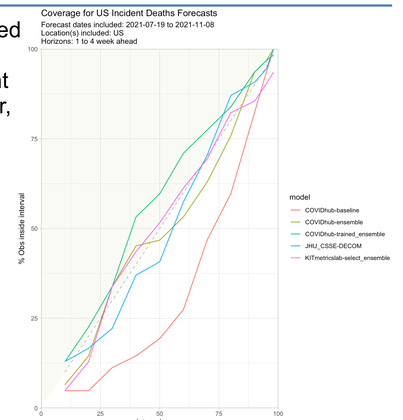
COVIDHub-ensemble national-level forecasts (blue) with 50%, 80% and 95% prediction intervals shown in shaded regions, and the ground-truth data (black) for incident cases (A), incident hospitalizations (B), incident deaths (C) and cumulative deaths (D). The truth data come from JHU CSSE (panels A, C, D) and HealthData.gov (panel B)



covidHubUtils: Forecast Evaluation

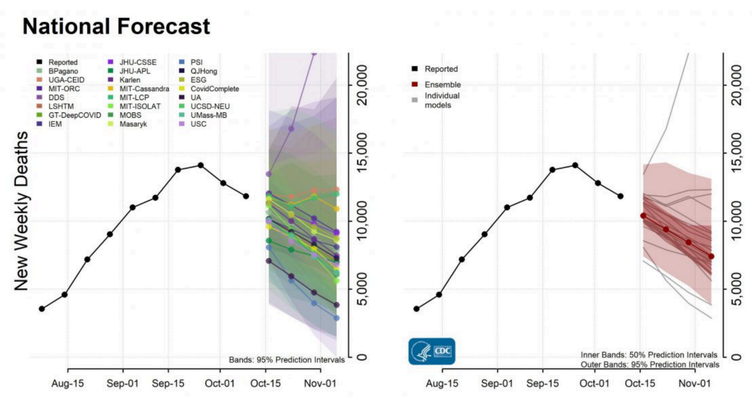
- **Bulk Query** for forecasts with selected locations, targets and horizons
- **Score forecasts** based on 7 different metrics including mean average error, coverage at varying prediction intervals and weighted interval score





Other Applications

Forecast data from the COVID-19 Forecast Hub is shared directly with the US CDC and is updated on an official COVID-19 Forecasting page hosted by the US CDC.



The structure of this dataset and our software infrastructure are also used in the European COVID-19 Forecast Hub and COVID-19 Scenario Modeling Hub. This large collaborative effort has provided data on short-term forecasts for over a year of forecasting efforts.

Key Links

- US COVID-19 Forecast Hub: https://github.com/reichlab/covid19-forecast-hub
- covidHubUtils source code: https://github.com/reichlab/covidHubUtils
- covidHubUtils vignettes: http://reichlab.io/covidHubUtils/

Takeaways

- It is important to establish a standardized format for forecasts data that represent a variety of modeling approaches, data sources and assumptions regarding the spread of COVID-19.
- The general structure of this data collection and software infrastructure could be applied to additional diseases or forecasting outcome in the future.
- Having a coordinated modeling group is efficient to inform policy and develop knowledge.

Acknowledgements: This material is based upon work supported by the U.S. Centers for Disease Control and Prevention under grant numbers 1U01IP001121 and by National Institutes of General Medical Sciences under grant numbers R35GM119582. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily represent the views of the U.S. Centers for Disease Control and Prevention, Food and Drug Administration, National Institutes of General Medical Sciences or the National Institutes of Health.The work of Johannes Bracher was supported by the Helmholtz Foundation via the SIMCARD Information and Data Science Pilot Project.