The COVID-19 Forecast Hub: using statistics and data science to support decision-making in a pandemic

Evan L. Ray

21 October 2020

https://covid19forecasthub.org/
COVID-19 Forecast Hub

**Team:** Martha Zorn, Nutcha Wattanachit, Serena Wang, Nicholas Reich, Evan Ray, Jarad Niemi, Khoa Le, Abdul Hannan Kanji, Dasuni Jayawardena, Katie House, Estee Cramer, Matt Cornell, Andrea Brennen, Johannes Bracher

* underline denotes ensemble contributor

**CDC Collaborators:** Michael Johansson, Matthew Biggerstaff, Joseph Walker, Velma Lopez, Rachel Slayton

**Ensemble "advisory committee":** Jacob Bien, Logan Brooks, Sebastian Funk, Tilmann Gneiting, Anja Muhlemann, Aaron Rumack, Ryan Tibshirani

**Modeling groups:** Over 50 groups at various institutions have contributed forecasts to the hub
Demo Visualization

https://viz.covid19forecasthub.org/
Why Forecast?

To inform public health planning:
• Do we have enough hospital beds?
• Where will we need to send more resources (personal protective equipment)?
• Where should we select vaccine trial sites?
Our Contributions

Infrastructure for collecting forecasts
• Each week, teams submit csv files with forecasts to GitHub repository
• Files are automatically validated for formatting and reasonableness
  • For example, predicted cumulative deaths can’t be negative

Models
• Ensemble model — combines forecasts from all submitted models
• Baseline model — naive reference model

Visualization
• Interactive visualizations shown earlier

Evaluation
• Are forecasts reliable?
• Are some models better than others?
Building the Ensemble: View 1

- For each combination of spatial unit $s$, time point $t$, and forecast horizon $h$, teams are required to submit $K=23$ quantiles of a predictive distribution:
  \[
  \hat{P}(Y \leq q_{s,t,h,1}^m) = 0.01, \quad \hat{P}(Y \leq q_{s,t,h,2}^m) = 0.025, \ldots, \quad \hat{P}(Y \leq q_{s,t,h,12}^m) = 0.5, \ldots, \quad \hat{P}(Y \leq q_{s,t,h,23}^m) = 0.99
  \]
  The predictive median
  Limits of a 98% prediction interval
- The predictive quantiles for the ensemble are a combination of component predictions at each quantile level:
  \[
  q_{s,t,h,k} = f(q_{s,t,h,k}^1, \ldots, q_{s,t,h,k}^M) \text{ for each } k = 1, \ldots, 23
  \]
Building an Ensemble — View 2

- The pairs \( \left( q_{s,t,h,k}^m, \hat{P}(Y_{s,t,h}^m \leq q_{s,t,h,k}^m) \right) \) fall along the predictive CDF for model \( m \)

- Three options for the combination function \( f \):
  - **QuantMean:**
    \[
    q_{s,t,h,k} = \frac{1}{M} \sum_{m=1}^{M} q_{s,t,h,k}^m
    \]
    Used through July 21, 2020
  - **QuantMedian:**
    \[
    q_{s,t,h,k} = \text{median}(q_{s,t,h,k}^1, \ldots, q_{s,t,h,k}^M)
    \]
    Used starting July 28, 2020
  - **QuantTrained:**
    \[
    q_{s,t,h,k} = \sum_{m=1}^{M} \beta_{t,h,k}^m \cdot q_{s,t,h,k}^m
    \]
    Evaluated, not released each week
Baseline Model

- Acknowledgment: idea adapted from a suggestion by Ryan Tibshirani
- Goal: Median predicted incidence is most recent observed incidence
- Predictions of cumulative deaths derived from predictions of incident deaths
Baseline Model

- Procedure:
  - Compute first differences of historical incidence:

\[ d_t = y_t - y_{t-1} \]

- Collect first differences and their negatives
- Sample first differences and add to last observed incidence (note: sampling not necessary for horizon 1, just use all observed differences)
- Adjustments for “niceness”:
  - Force median = last observed incidence
  - Truncate at 0
- Iterate for horizons > 1
Evaluation Part 1: Ensembles Compared (WIS)

- Weighted Interval Score (WIS) measures the distance of the predictive distribution from the observed response
  - Sum of mean absolute error and penalties for predictive intervals that miss
  - Smaller WIS is better

- All three ensembles are better than the baseline
- QuantMedian and QuantTrained are comparable to or better than QuantMean
- No clear ordering of QuantMedian and QuantTrained
Evaluation Part 2(b): Ensemble vs Components (WIS)

Figure due to Nick Reich
Evaluation Part 2(a): Ensemble vs Components (MAE)

- MAE: On average, how far was the median of the predictive distribution from the eventually-observed count?
- Looking here at results for a set of 8 models that have submitted forecasts for all states and the US since the week of May 2
- Credit to Estee Cramer for this figure
Evaluation: Probabilistic Calibration

- At each predictive interval level, what proportion of intervals contain the eventually-observed outcome?
- This figure shows calibration of the ensemble model only.
Evaluation: Probabilistic Calibration

- At each predictive interval level, what proportion of intervals contain the eventually-observed outcome?
- This figure shows calibration of the ensemble model only

Of the 120 forecasts where rounding changed coverage status for 95% intervals, 99 were for weeks with no new deaths
A Current Challenge: Case Forecasts
Current Work/Future Directions

• Continue accepting and processing forecasts
  • Challenges:
    • Massive amount of data — how to store it?
    • Versioning forecasts

• Refining ensemble methods
  • Recent analyses indicate a weighted median may be helpful
  • Challenges:
    • Models change over time; relative skill may not be stable
    • Limited data, flexible ensemble approaches may be overfitting

• Forecasting model development
  • In flu forecasting, we saw that statistical time-series models were quite effective — we would like to develop these models for COVID-19
  • Challenges:
    • Time
    • Hierarchical structure, forecast horizon-specific model fits, merging mechanistic models with time series models
Thanks!

Acknowledgments again to:
- Nick Reich
- The whole COVIDhub team
- CDC colleagues
- Contributing modeling teams
- Epidemiologists and medical workers everywhere

We’re hiring for a post-doctoral position, get in touch if interested!
elray@umass.edu