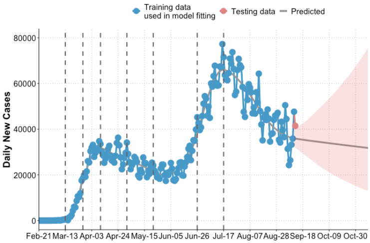
Survival-Convolution Models for Predicting COVID-19 Pandemic and Assessing Effects of Mitigation Strategies

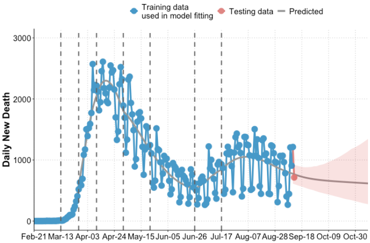
Yuanjia Wang

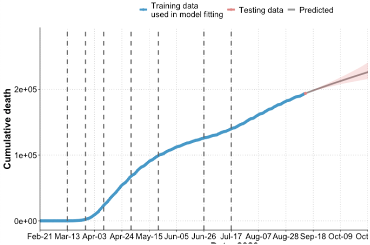
Department of Biostatistics, Mailman School of Public Health, Columbia University

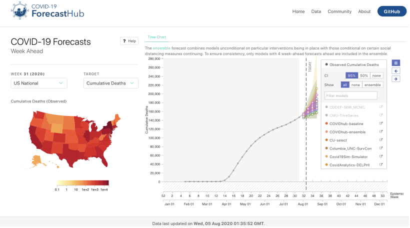
& Division of Biostatistics, New York State Psychiatric Institute

Daily Forecasts of COVID-19 Pandemic





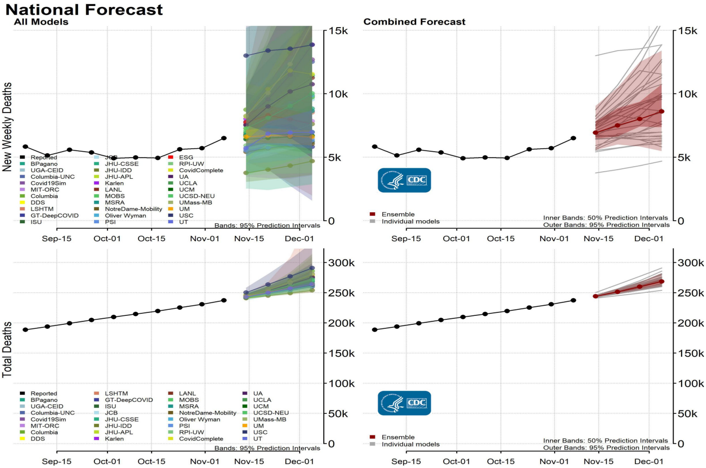




We submit our forecasts to [COVID Forecast Hub](), which is used by the US Centers for Disease Control and Prevention ([CDC]())

CDC Forecasts

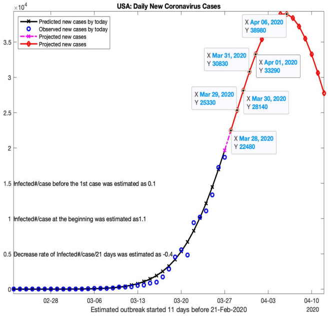
The ensemble forecast predicts that 260,000 to 282,000 total COVID-19 deaths will be reported by December 5[2].



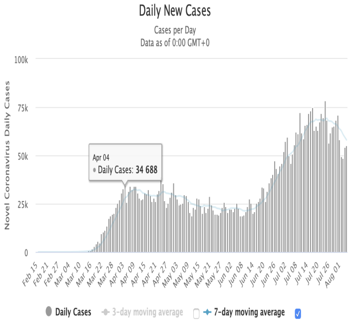
[2]CDC Forecasts: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html

Starting Points

First patient reported in NYC on March 1. Stay-at-home-order issued on March 22. Unprecedented response measures: lockdown, travel restrictions, social distancing, closure of schools, businesses.

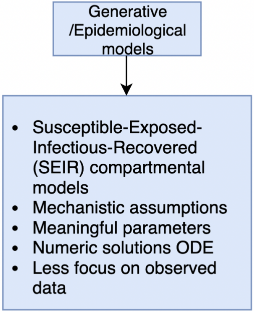


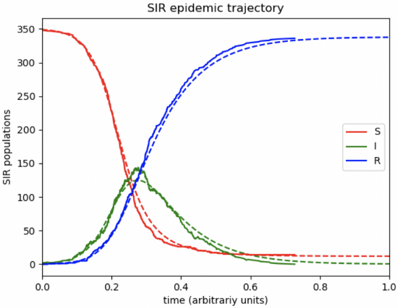
Daily New Cases in the United States



Methods

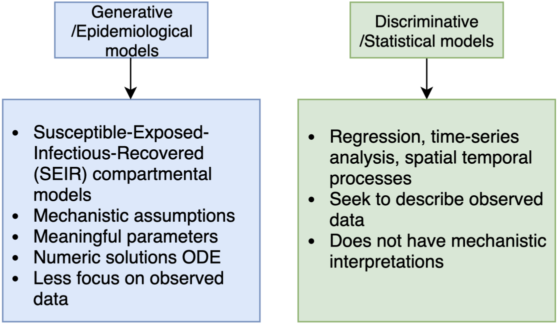
Figure. Epidemiological Compartmental Models
(e.g., Susceptible-Infectious-Recovered; SIR)





Infectious Disease Models

Figure. Epidemiological Models and Statistical Models



Goal 1: Combine nonparametric curve fitting with mechanistic-based SEIR model (provide important parameters, i.e., effective reproduction number R_t).

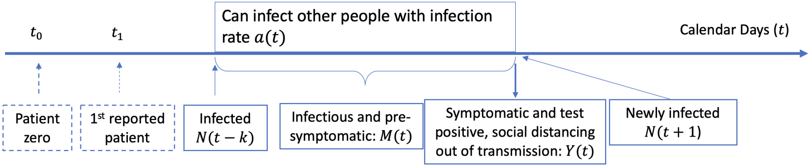
Goal 2: Natural experiment to evaluate mitigation strategies. SEIR models rely on a large number of unknown parameters.

Modeling Considerations

- What is the forecast target? Peak week, size, duration, cumulative/incident cases.
  Predict daily incident cases and incident deaths at the national-level and state-level.
- Important factors for modeling:
  - SARS-CoV-2 virus has a long incubation period (up to 14 days, extreme case 21 days)
  - Highly infectious in the pre-symptomatic phase: 50% transmission occurred during this phase (US CDC)
  - Time-varying transmission rate as public health interventions are implemented and societal behavior changes
  - Intervention effect may be time-dependent
- Transparency, robustness are important for policy decision making.

Survival-Convolution Model

- $M(t) = \sum_{k=0}^{\infty} N(t-k)S(k)$
- $Y(t) = \sum_{k=0}^{\infty} N(t-k)[S(k) - S(k+1)]$
- $N(t+1) = a(t)[M(t) - Y(t)]$



- $N(t)$ number of new infections on date t.
- At time t, number of the patients who have been infected for k days and remain in the transmission chain (e.g., pre-symptomatic):

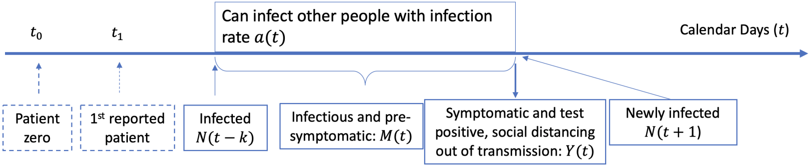
$$N(t-k)S(k),$$

  $S(k)$ proportion of infected persons remaining infectious and in the transmission chain after k days of exposure (discrete survival function of time to out of transmission).
- Total number of infectious persons right before t (including pre-symptomatics):

$$M(t) = \sum_{k=1}^{C} N(t-k)S(k).$$

Survival-Convolution Model

- $M(t) = \sum_{k=0}^{\infty} N(t-k)S(k)$
- $Y(t) = \sum_{k=0}^{\infty} N(t-k)[S(k) - S(k+1)]$
- $N(t+1) = a(t)[M(t) - Y(t)]$



- Total number of cases out of transmission on date t:

$$Y(t) = \sum_{k=1}^{C} N(t-k)[S(k) - S(k+1)].$$

- Denote the effective transmission rate by $a(t)$,

$$N(t) = a(t)(M(t) - Y(t))$$

$$N(t) = a(t) \sum_{k=1}^{C} N(t-k)S(k+1). \tag{1}$$

Equation (1) gives a convolution update for the number of new infections given the past infections $N(t-1), N(t-2), \ldots, N(t_0)$.

Modeling Transmission Rate

Model $a(t)$ as non-negative, piece-wise linear functions with knots placed at meaningful event times:

- Before report of first case t_1, transmission rate is a constant a_0.
- Once the first positive case was reported, the society starts to respond, so model the transmission rate with a linear function.
- When a massive public health intervention (e.g., nation-wide lockdown) is implemented, introduce an additional linear function with a new slope parameter.
- The simplest model has only 2 parameters (a_0, a_1)!

Time-varying Effective Reproduction Number

Effective reproduction number (R_t): the average number of secondary cases infected by primary cases who are infectious at time t[3]

$$R_t = \frac{N(t)}{\sum_{k=1}^{C} N(t-k)w(k)}$$

$w(k)$ probability mass function of the serial interval distribution between primary and secondary cases (Gamma distribution with shape and scale parameters $(4.36, 1.10)$[4]).

R_t captures the temporal changes in the disease spread.

[3] Cori, A., Ferguson, N. M., Fraser, C., Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. American Journal of Epidemiology, 178(9), 1505-1512.

[4] Nishiura, H., Linton, N. M., Akhmetzhanov, A. R. (2020). Serial interval of novel coronavirus (COVID-19) infections. International Journal of Infectious Diseases.

Evaluation of Public Health Intervention Effect

Quasi-experiments longitudinal pre-post intervention design. Often used to study health policies when randomized trials are not feasible.

Assumptions:

- Local randomization: subjects infected before or after intervention are similar within a short period of time
- Continuity: the trend before implementation continues had the intervention not been implemented

Intervention effects estimated as the difference in the slope of $a(t)$ before and after an intervention takes place. Corrects for the natural decline of the transmission rate over time.

Estimation Using Confirmed Daily Cases

Let θ denote all parameters in the infection rate $a(t)$ and t_0.

Let $Y_o(t_1), Y_o(t_1+1), Y_o(t_1+2), ..., Y_o(t_n)$, denote the daily new COVID-19 cases reported from t_1 to the last date t_n in the training set.

Model observed cases accounting for measurement errors:

$$Y_o(t_i) = Y(t_i;\theta) + \sqrt{Y(t_i;\theta)}\epsilon(t_i),$$

$\epsilon_i(t)$ is a normalized residual error (e.g., reporting error), with variability proportional to $Y(t;\theta)$.

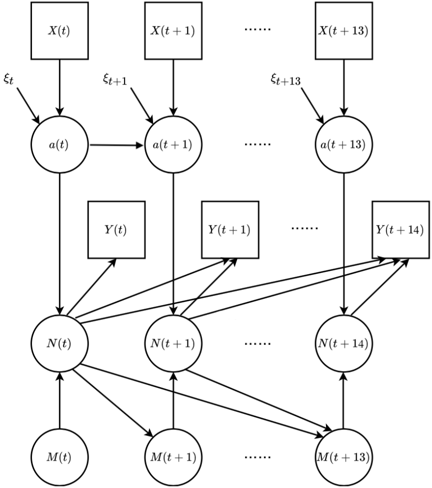


Figure A1: Illustration diagram of the spatial-temporal model for one consecutive 14 days (the maximum incubation period of COVID-19). $M(t)$: number of infected subjects who remain in the transmission chain and can transmit virus to others (including those who are pre-symptomatic or asymptomatic) on day t. $N(t)$: number of newly infected subjects on day t. $Y(t)$: number of diagnosed subjects out of transmission chain on day t. $a(t)$: infection rate on day t, which depends on area characteristics $X(t)$ and spatial-temporal transmission model parameters ξ_t.

Optimization and Inference

Optimization:

- Objective function: squared error on the predicted number of cases and the observed after a square-root transformation:

$$\sum_{t_1 \le t \le t_n} \left[\sqrt{Y_o(t)} - \sqrt{Y(t;\theta)} \right]^2$$

- Stochastic gradient descent implemented in Tensorflow.

Inference:

- Assume that the standardized residuals are exchangeable.

$$[Y_o(t) - Y(t;\theta)] / \sqrt{Y(t;\theta)}$$

- Permutation of predicted standardized residuals over time $\tilde{\epsilon}(t) = \left[Y_o(t) - Y(t;\widehat{\theta}) \right] / \sqrt{Y(t;\widehat{\theta})}$.
- Generate new copies of daily cases, $\tilde{Y}(t) = Y(t;\widehat{\theta}) + \sqrt{Y(t;\widehat{\theta})} * \tilde{\epsilon}(t)$ and repeat permutation N times.

Forecast Daily Incident Deaths

Let $Z(t)$ denote number of incidence deaths at day t, convolution

$$Z(t) = b(t) \sum_{k=0}^{C_2} N(t-k) P(T_1 + T_2 = k),$$

$b(t)$ case fatality rate, T_1 time from initially infected to symptomatic, T_2 time from symptomatic to death.

Optimization: combine loss function of cases and deaths.

Inference: jointly permute residuals from the incident case and incident death model.

For forecasts, extrapolate current estimated parameters on $a(t)$.

Analysis Details

Data

Numbers of daily confirmed new cases and new deaths can be obtained from many public sources.

- National level: a publicly available database that curates and validates multiple sources on COVID-19 statistics `www.worldometers.info/coronavirus`
- State level: JHU Center for System Science and Engineering (CSSE) `https://github.com/CSSEGISandData/COVID-19`

Model Setups

Countries to analyze: China, South Korea, Italy, US

China and South Korea: a single piece for $a(t)$. About two weeks data for training and the rest of data up to May 10 as testing data. Infection rate:

$$a(t) = \begin{cases} a_0^+ & t < t_1 \\ (a_0 + a_1(t - t_1))^+ & t \geq t_1 \end{cases}$$

3 parameters: t_0, a_0, a_1.

Goal: examine prediction performance.

Model Setups

Italy: 4 pieces. A knot at nation-wide lockdown (March 11, t_2), two knots with two weeks apart afterwards (March 25, t_3; April 8, t_4). Capture the immediate, short-term and mid-term intervention effect (March 25, t_3; April 8, t_4).

$$a(t) = \begin{cases} a_0^+ & t < t_1, \\ (a_0 + a_1(t - t_1))^+ & t_1 \leq t < t_2, \\ (a_0 + a_1(t_2 - t_1) + a_2(t - t_2))^+ & t_2 \leq t < t_3, \\ (a_0 + a_1(t_2 - t_1) + a_2(t_3 - t_2) + a_3(t - t_3))^+ & t_3 \leq t < t_4 \\ (a_0 + a_1(t_2 - t_1) + a_2(t_3 - t_2) + a_3(t - t_3) + a_4(t - t_4))^+ & t \geq t_4. \end{cases}$$

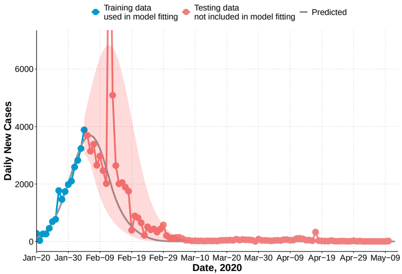
Goal: estimate lockdown effect, i.e., immediate effect (a_2 vs a_1), short-term (a_3 vs a_1), midterm (a_4 vs a_1).

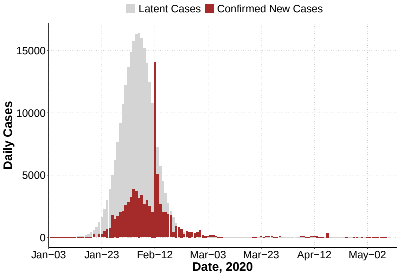
US (10-20 days behind Italy): 5 pieces. A knot at the declaration of national emergency (March 13, t_2) and 3 knots (2-week apart) afterwards (March 27, April 10, April 24).

National-level Analysis Results

China

Training data: January 20 to February 4; testing data: February 5 to May 10.

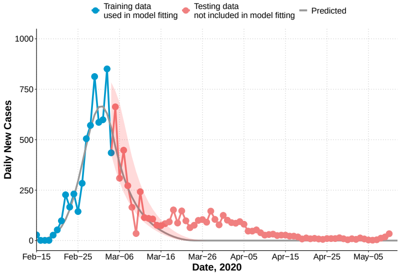


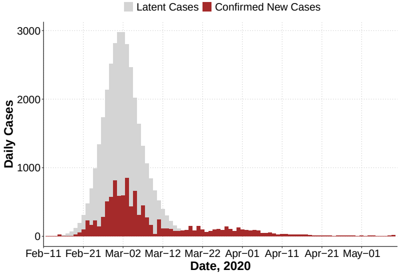


- t_0: Jan 3 (17 days before first report)
- Predicted total: 58,415; 95% CI: (42,516, 133,083)
- Observed total: 82,901. Two outliers on Feb 12, 13. Excluding outliers: 62,356.

South Korea

Training data: February 15 to March 4; testing data: March 4 to May 10.

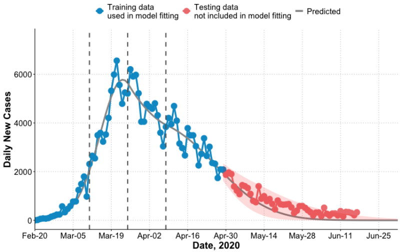


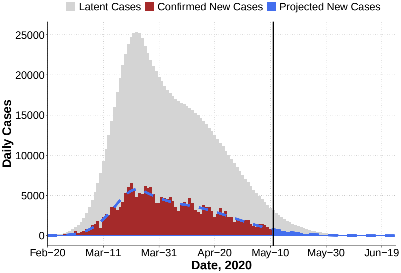


- t_0: Feb 11 (4 days before first report)
- Small outbreak after March 15 not captured
- Predicted total by March 15: 7,816
- Observed total: 8,162.

Italy

Training data: February 20 to April 29 (7 weeks after lockdown); testing data: April 30 to June 15.

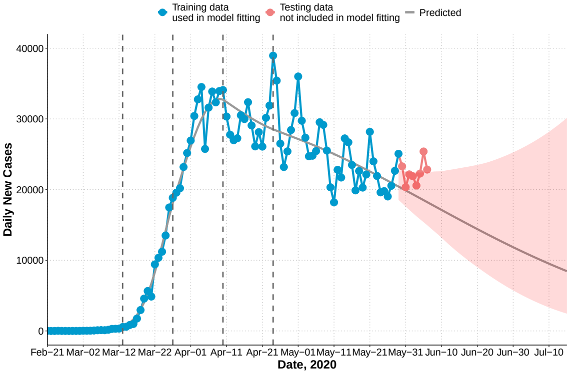




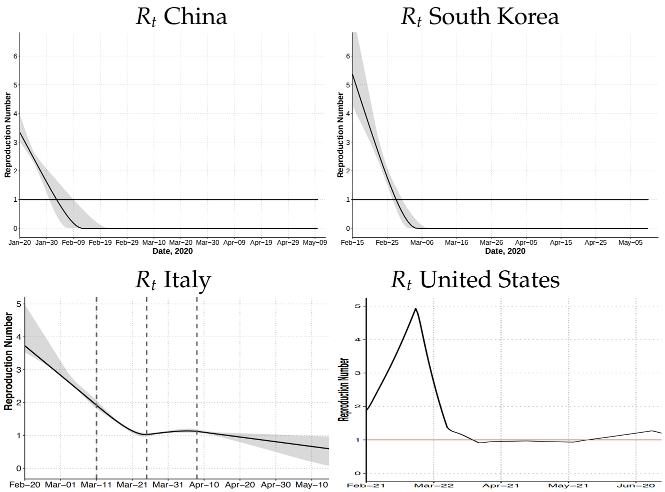
- t_0: Feb 10 (10 days before first report)
- Predicted total by May 31: 223,079 (CI: 202,940, 263,152)
- Observed total: 232,997
- Rate of decrease after the peak is slower than rising (asymmetric)

US

Training data: February 21 to May 29. One knot on March 13 and 3 knots two-weeks apart (3/27, 4/10, 4/24).



Had the late spring trend continue, total cases: 2.7 million, total deaths: 157K. Date with < 100 cases: Nov 9. But already observed an uptick late May.



- R_t reduced to < 1.0 in 2 weeks in China and South Korea. R_t reduced to 1.0 in 6 weeks in Italy (remained around 1.0 for 3 weeks). Nation-wide lockdown in Italy did not significantly further reduce the rate of decrease ($p > 0.05$). US R_t reduced to < 1 in 7 weeks, flat for 6 weeks before increasing again.

Comparing Infection Rates $a(t)$

Country	Parameter	Estimate	95% CI
China	a_0	0.793	(0.68, 1.02)
	a_1	-0.693	(-1.13, -0.42)
	Duration	44	(39, 55)
South Korea	a_0	1.363	(1.03, 1.98)
	a_1	-1.496	(-2.39, -0.96)
	Duration	39	(37, 43)
Italy	a_0	0.789	(0.73, 1.10)
	a_1	-0.358	(-0.68, -0.26)
	a_2	-0.372	(-0.46, -0.31)
	a_3	0.061	(0.02, 0.12)
	a_4	-0.057	(-0.12, -0.01)
	Duration	123	(103, 179)
United States	a_0	0.774	(0.73, 0.78)
	a_1	-0.029	(-0.03, 0.03)
	a_2	-0.665	(-0.69, -0.54)
	a_3	-0.173	(-0.23, -0.13)
	a_4	0.018	(-0.01, 0.05)
	a_5	-0.005	(-0.02, 0.01)
Continue current†	Duration	262	(187, ∞)

National-level Forecasts

First time the model predicted a third surge: August 21, 2020. (Left: incident cases; Right: incident deaths).

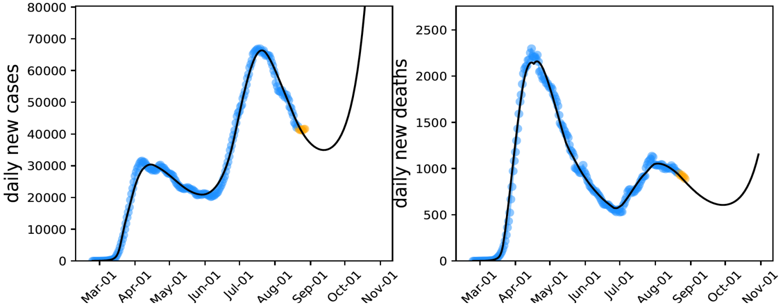
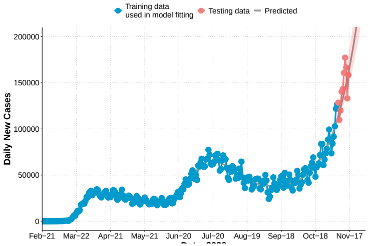
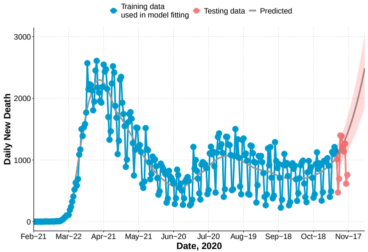
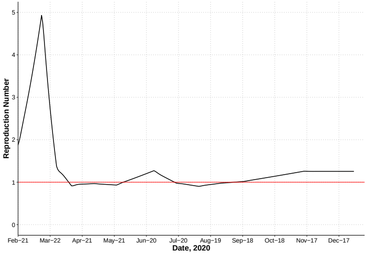


Figure. Current forecasts of incident cases, incident deaths and R_t. Training data up to November 7 (week 46).







Without change in $a(t)$,

- Incident cases will reach 200k on 11/22 (actual date: 11/20).
- Cumulative deaths will reach 300k on 12/14 (actual date: 12/14).
- $R_t = 1.25$.

Model using training data up to 12/24/2020 predicts cumulative deaths reaching 400k by 1/18/2021. New variant a huge concern.

Performance of Models During Summer Surge: July, 2020

Figure. Independent Evaluation by CovidComplete

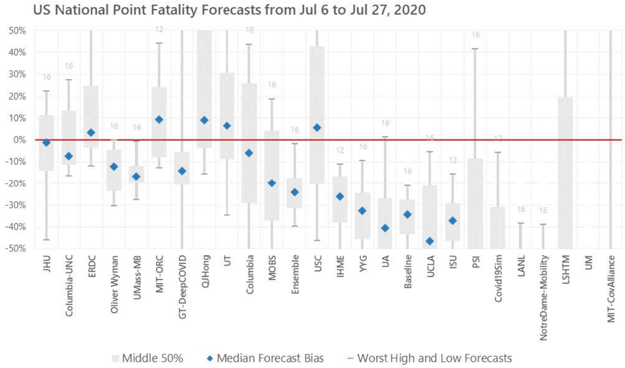


Figure. Example of Performance on Week 33 (7/12/2020)

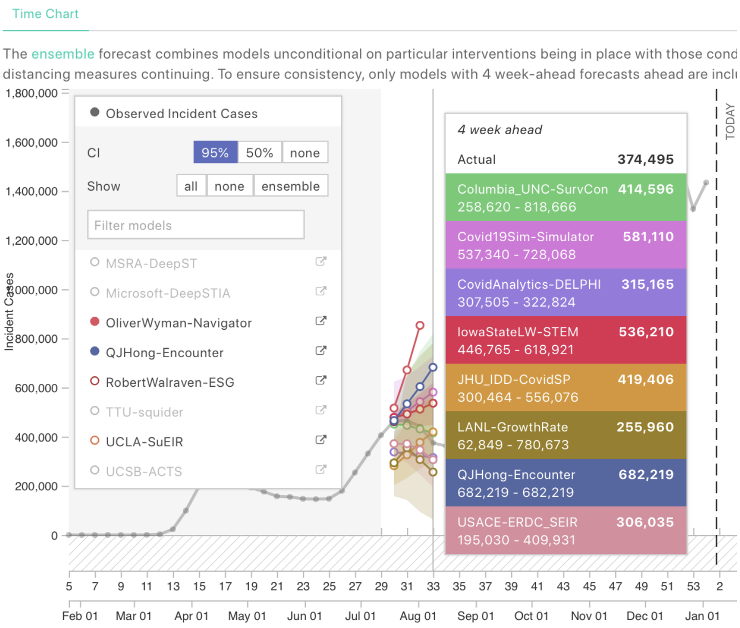
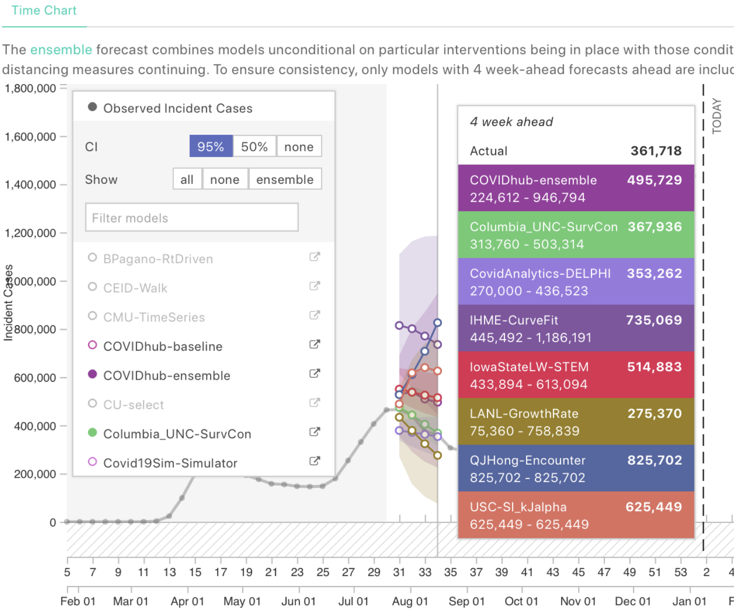
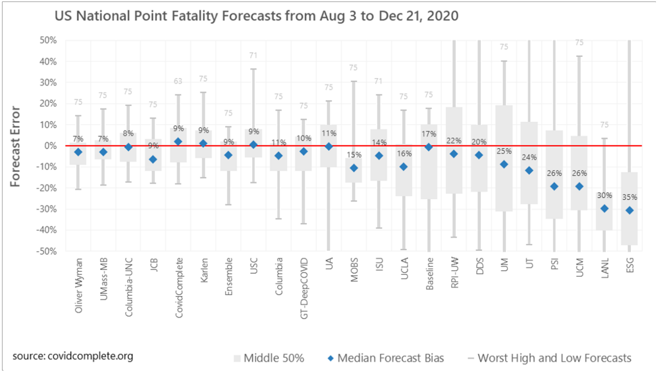


Figure. Example of Performance on Week 34 (7/19/2020)

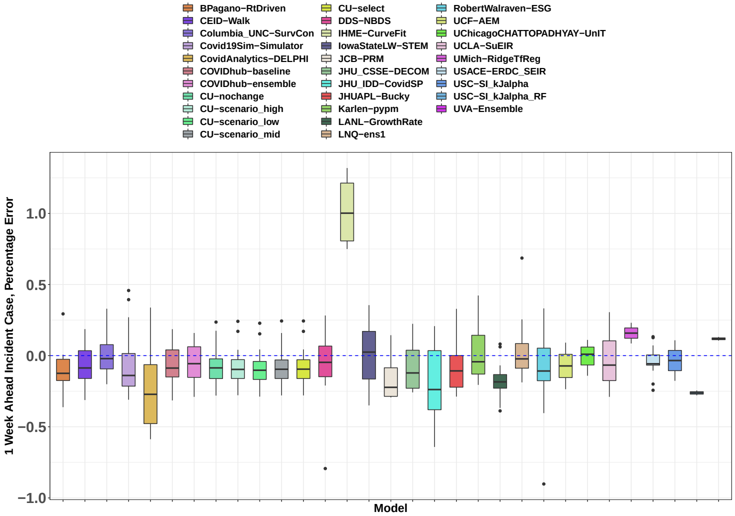


Performance of Models Post July: 8/3/2020 –12/21/2020

Figure. Independent Evaluation by CovidComplete

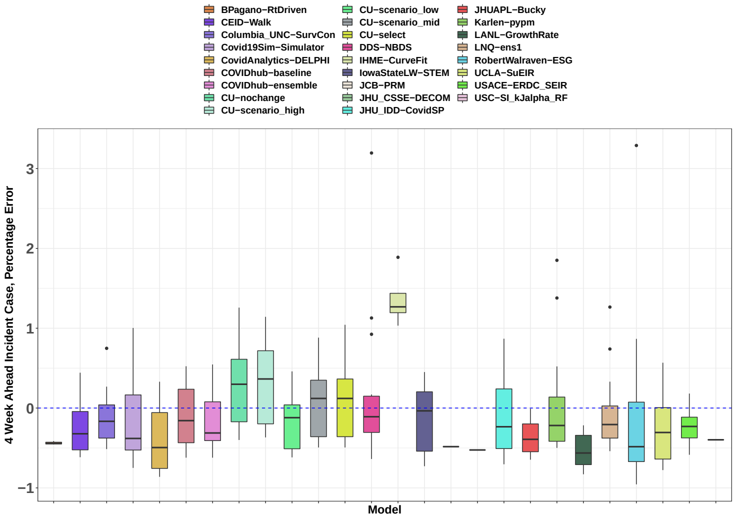


Performance of Cases Model 7/5/2020-1/2/2021



Yuanjia Wang, Department of Biostatistics, Columbia University

Performance of Cases Model 7/5/2020-1/2/2021



Coverage Probability

Table. Coverage Probability of 95% Prediction Intervals of Forecasts (since July 5, 2020)

Model	1 Week	2 Week	3 Week	4 Week
COVIDhub-ensemble	0.813	0.933	1.000	1.000
Columbia-UNC-SurvCon	0.938	0.933	1.000	0.923
GT-DeepCOVID	0.938	0.933	0.857	0.846
IowaStateLW-STEM	0.533	0.357	0.538	0.500
NotreDame mobility	0.250	0.267	0.286	0.308
CovidAnalytics DELPHI	0.750	0.600	0.500	0.462
IHME CurveFit	0.455	0.600	0.700	0.800

Discussion

Summary

Propose a survival convolution model for forecast daily incident cases, deaths, estimate R_t, and comparison of mitigation strategies.

Simpler hybrid statistical/epidemiological models can be useful and robust for population science (full SEIR models require careful calibration; may work better with individual level data).

Challenges: difficult to make long term forecast

- Incomplete knowledge on the drivers of the epidemic
- Lack of data on behavioral change and policy enforcement; difficult to predict societal behavioral change
- Lack of accurate data on cases and deaths (reporting delay, limited testing capacity)

References and Acknowledgements

- Wang Q et al. (2020). Survival-convolution models for predicting COVID-19 cases and assessing effects of mitigation strategies. *Frontiers in public health* 8: 325. Github: https://github.com/COVID19BIOSTAT/covid19_prediction
- Chen Y et al. (2021). Dynamic COVID risk assessment accounting for community virus exposure from a spatial-temporal transmission model. *NeurIPS*, 34.
- Xie S et al. (2022). Evaluating Public Health Intervention Strategies for Mitigating COVID-19 Pandemic. *Statistics in Medicine*. In press. https://doi.org/10.1002/sim.9482
- COVID-19 Forecast Hub Consortium (2022). *PNAS* 119 (15), e2113561119.

- Funding agency: GM124104A1-S1.

Thank You!