
Further Supplementary Information

A prospective evaluation of AI-enabled epidemiology to forecast COVID-19 in the USA and Japan

The aim of this supplementary information is to provide further information to support the claims made in the Article "*A prospective evaluation of AI-enabled epidemiology to forecast COVID-19 in the USA and Japan*". It is the hope of the authors that by providing these supplementary results and associated discussion that the conclusions of the letter are strengthened, along with the reproducibility of the work.

We present the following supplementary material:

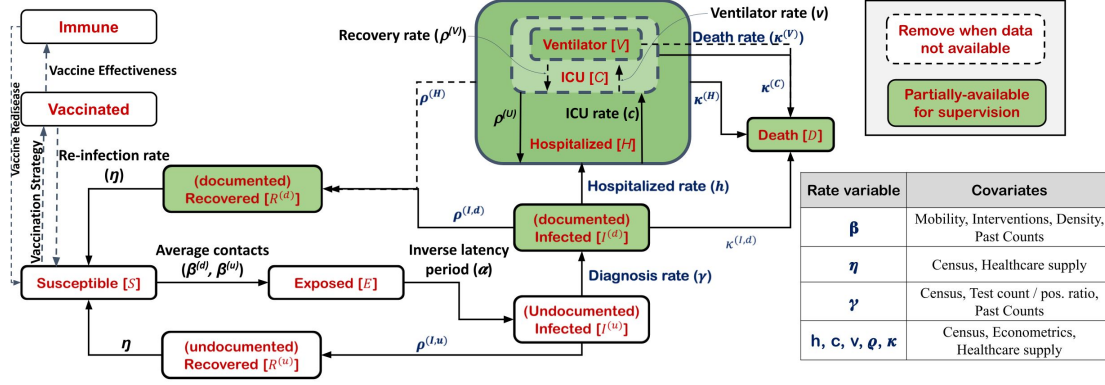
- Supplement 1 provides detail on the proposed model and the changes implemented during the prospective period.
- Supplement 2 provides supplementary methods and evaluation details, including full reporting of performance for individual locations and model comparison with baselines.
- Supplement 3 provides additional results from counterfactual simulation experiments in the USA and Japan.
- Supplement 4 provides model explainability graphs for each US State and Japanese prefecture, and detailed results on feature ranking from the encoders.
- Supplement 5 provides results from ablation studies.
- Supplement 6 provides the full results of the model fairness analysis.
- Supplement 7 provides an analysis of and justification for using model uncertainty to reject low-confidence forecasts.
- Supplement 8 provides anecdotal case studies from organisations who used the model in practice during the prospective study.

1 Proposed compartmental model

Figure 1 overviews our compartmental model. The transitions between the compartments are formulated in the Methods section.

Major model changes during the prospective evaluation period

The prospective trial began after 28 day predictions were publicly launched in the US and Japan. As the models were public and being used by policy makers, some elements of the model were altered during the trial to ensure the model was stable and accurate. This maintenance work was essential to address the changing requirements of the model, such as changing data sources and to improve stability.



Supplementary Figure 1 | Framework architecture and modeled compartments. Modeled compartments and the corresponding features are shown, with the legend on the right. Each rectangle represents a compartment, while each connection between rectangles represents a transition between compartments. The rates are named in the legend. Rectangles shaded in green represent compartments that are partially supervised.

The changes reflect the real world experience of deploying and maintaining an AI model, especially in a rapidly-changing pandemic environment. All changes only influence performance after they were implemented; the prospective performance we report for each date is based on the prediction of our model made on that date.

- **11/11/2020:** 28 day forecast model was launched in the USA and Japan.
- **11/20/2020:** Testing related features were converted to per capita quantities from raw counts.
- **12/06/2020:** (i) Additional features for antibody/antigen counts and weather data were added. (ii) The earlier version of the model was using the predicted features instead of actual ones during training, it was switched to actual ones so that the encoder weights can learn how to better work with the true features.
- **12/09/2020:** (i) Number of iterations for quantile regression was increased. (ii) The partial teacher forcing coefficient for the Japan model was reduced for undocumented infected as it was overpredicting the new confirmed cases due to overpredicted diagnosis rate.
- **12/23/2020:** (i) Directional penalty regularization was added for mobility and NPI features for US State model. (ii) The cutoff for R_{eff} regularization was converted to a tunable hyperparameter. (iii) Search symptoms data was integrated as a feature. (iv) Quantile regression training loss coefficients were re-optimized to get better validation WIS values. (v) Random sampling of training range was introduced (See Algorithm 1) to reduce train-test mismatch.
- **12/31/2020:** (i) Quantile normalization was switched from mean to median, to ensure the mid-quantile better matches the point forecasts.

-
- **01/08/2021:** (i) Directional penalty regularization was added for US County and Japan models. (ii) Quantile regression training loss coefficients were reoptimized to get better validation WIS values for the Japan model.
 - **02/26/2021:** Two-dose vaccine modeling was introduced reflecting the most recent experimental results[1] (see Methods) to be used in baseline forecasts and counterfactual analysis.

2 Performance Evaluations

Prospective performance for all US states and Japanese prefectures

We provide the 4-week predictive performance in APE (see Methods) of both the USA and Japan model at multiple time points during the prospective trial. Performance is shown for forecasts of deaths (Supplementary Table 1) and cases (supplementary Table 2)) for all 51 US states. For Japan, performance is shown for deaths (Supplementary Table 3) and cases (supplementary Table 4) all 47 Japan prefectures.

Supplementary Table 1 | MAPE (%) for 4-week forecasts of deaths for all the US states on different prediction dates.

Forecast Date \Rightarrow State Name \Downarrow	2020-11-15	2020-11-22	2020-11-29	2020-12-06	2020-12-13	2020-12-20	2020-12-27	2021-01-03	2021-01-10
Alabama	4.6	2.1	9.2	3.6	8.8	9.5	7.6	17.8	23.0
Alaska	1.6	37.4	38.8	22.0	25.8	27.1	1.7	9.0	5.0
Arizona	0.5	4.9	6.3	5.1	6.1	6.7	7.9	8.1	3.1
Arkansas	2.1	1.2	5.0	11.7	1.9	6.0	1.3	5.3	10.7
California	1.7	6.0	8.1	5.4	5.5	4.1	10.5	8.4	0.8
Colorado	8.2	5.1	3.8	3.1	8.2	16.1	7.3	7.1	12.5
Connecticut	0.0	2.1	1.4	2.0	3.8	2.1	0.4	1.9	5.3
Delaware	1.7	5.8	6.8	4.7	6.1	8.8	8.3	1.4	3.3
District Of Columbia	0.9	4.4	2.6	1.0	1.9	0.7	0.1	0.8	1.1
Florida	6.5	1.0	0.8	2.6	3.6	0.6	1.4	0.4	2.4
Georgia	15.8	0.7	1.1	3.0	4.7	0.1	4.8	8.5	9.3
Hawaii	7.7	7.7	9.9	5.0	7.6	1.0	5.9	18.5	9.2
Idaho	1.3	3.3	0.6	3.3	12.5	13.4	7.3	2.6	7.9
Illinois	3.3	2.1	1.6	4.8	2.0	0.2	1.8	2.4	6.5
Indiana	0.8	3.7	0.5	1.3	0.3	0.7	0.1	5.4	2.6
Iowa	4.7	4.2	13.2	15.3	5.5	3.0	1.9	4.0	4.1
Kansas	16.8	14.0	10.2	8.6	8.4	2.3	1.4	9.6	14.2
Kentucky	0.0	2.9	1.8	1.2	1.8	1.1	1.8	7.9	6.8
Louisiana	2.5	1.5	0.9	2.5	1.7	0.0	2.1	0.5	4.4
Maine	1.1	6.8	7.1	10.7	20.0	0.4	8.5	4.7	13.6
Maryland	4.6	2.8	1.4	0.2	1.2	2.4	1.9	1.3	4.8
Massachusetts	2.2	1.3	0.6	1.2	2.8	3.8	2.0	1.2	4.4
Michigan	4.1	0.1	3.7	3.3	3.8	1.7	1.2	1.7	1.8
Minnesota	8.3	9.9	1.6	5.2	2.7	2.5	2.4	0.2	3.7
Mississippi	2.9	2.7	0.6	2.7	0.7	1.4	3.5	1.8	0.5
Missouri	2.5	7.7	14.8	8.6	7.9	0.2	8.4	3.3	1.5
Montana	10.3	10.4	8.6	1.8	3.2	8.0	10.2	10.4	2.6
Nebraska	5.8	12.1	0.1	2.9	19.5	7.5	0.3	4.6	10.5
Nevada	5.0	5.2	3.6	2.2	4.5	1.0	3.8	7.7	2.1
New Hampshire	1.2	3.7	1.3	4.7	0.6	7.2	1.8	7.6	9.8
New Jersey	0.2	1.0	0.1	0.4	0.4	0.8	1.7	0.8	3.1
New Mexico	3.8	11.1	18.9	8.0	1.6	0.9	2.8	0.1	4.4
New York	3.2	2.1	2.5	2.9	2.6	2.1	0.8	2.7	2.9
North Carolina	0.5	1.8	1.0	1.1	2.4	0.0	3.1	3.9	1.1
North Dakota	15.0	4.4	2.0	6.8	6.0	4.3	4.8	2.2	1.3
Ohio	3.3	5.0	5.2	5.6	4.4	0.9	3.5	1.8	1.5
Oklahoma	3.5	4.9	0.3	1.7	0.2	3.4	8.9	7.0	2.3
Oregon	13.1	12.3	3.3	4.4	7.1	9.9	10.6	5.2	1.0
Pennsylvania	9.9	3.4	4.1	2.2	0.4	0.4	4.6	3.4	0.4
Rhode Island	3.5	3.3	1.4	1.4	7.2	10.6	10.0	1.0	12.4
South Carolina	3.5	0.9	2.5	1.9	0.3	2.3	4.0	3.7	1.5
South Dakota	10.1	1.4	1.9	3.8	7.5	0.7	2.8	1.3	0.4
Tennessee	1.0	1.5	5.3	6.7	1.6	4.8	1.5	6.3	1.7
Texas	2.2	0.5	1.2	1.5	3.7	4.8	4.8	5.4	1.3
Utah	12.1	16.6	22.4	29.0	28.4	2.0	5.4	7.3	6.1
Vermont	3.7	11.1	17.4	11.8	6.1	13.9	4.5	12.9	41.0
Virginia	5.5	1.0	1.1	3.2	1.1	1.4	0.1	2.4	1.8
Washington	11.1	8.7	9.2	11.1	3.4	8.9	11.5	2.4	12.9
West Virginia	6.3	10.8	2.0	0.0	2.8	2.6	2.1	4.0	11.7
Wisconsin	12.8	7.4	5.0	3.8	1.4	2.3	1.1	4.5	0.5
Wyoming	20.0	25.4	14.1	5.0	7.6	11.4	21.6	9.7	6.5

Supplementary Table 2 | MAPE (%) for 4-week forecasts of confirmed cases for all the US states on different prediction dates.

Forecast Date \Rightarrow State Name \Downarrow	2020-11-15	2020-11-22	2020-11-29	2020-12-06	2020-12-13	2020-12-20	2020-12-27	2021-01-03	2021-01-10
Alabama	10.4	8.3	5.9	6.5	5.2	14.9	2.9	0.8	8.0
Alaska	1.5	12.7	18.5	12.9	23.7	14.8	2.7	9.9	8.4
Arizona	14.6	6.7	4.4	3.7	4.9	5.3	11.8	6.5	20.7
Arkansas	5.4	2.1	2.3	4.9	1.7	3.3	4.7	0.1	13.8
California	22.1	17.1	16.6	5.9	22.5	31.0	8.8	2.5	7.7
Colorado	7.6	30.7	26.2	9.1	17.0	10.0	1.6	3.8	13.4
Connecticut	11.2	0.2	0.4	1.1	4.0	14.0	32.4	19.0	37.5
Delaware	14.3	7.5	1.4	2.8	4.7	15.6	5.9	2.9	16.5
District Of Columbia	14.2	7.2	1.8	7.4	4.2	11.4	5.7	6.5	2.5
Florida	12.8	0.5	0.3	1.6	1.4	2.4	2.0	5.6	22.1
Georgia	4.9	13.2	10.7	8.9	2.1	0.1	4.5	0.3	3.7
Hawaii	3.0	9.1	9.1	2.2	1.7	1.1	4.7	3.8	14.9
Idaho	5.5	7.7	5.7	0.5	10.4	9.6	18.0	5.9	11.7
Illinois	1.3	32.3	14.3	0.9	3.2	6.5	4.2	0.9	12.4
Indiana	5.0	18.5	11.5	0.1	10.5	11.5	2.4	3.8	14.9
Iowa	12.8	33.9	13.5	1.8	1.3	0.1	3.0	2.7	11.1
Kansas	10.4	7.2	14.9	7.9	2.9	4.4	13.9	13.5	24.0
Kentucky	10.9	8.0	10.0	5.1	13.5	3.4	5.8	5.7	17.8
Louisiana	13.7	6.1	0.9	13.7	6.4	2.9	8.0	5.1	14.3
Maine	1.3	6.4	18.5	8.7	22.0	13.5	0.7	12.3	38.7
Maryland	9.3	2.3	3.8	2.9	2.5	1.3	4.9	5.6	12.5
Massachusetts	22.1	5.6	3.1	9.7	1.3	25.4	1.4	3.1	8.5
Michigan	9.9	37.9	35.6	19.3	2.5	6.9	8.0	6.9	10.9
Minnesota	1.8	41.9	39.7	12.6	10.4	6.9	0.9	2.1	8.0
Mississippi	10.3	4.0	7.1	2.8	8.2	4.1	4.5	0.6	11.6
Missouri	7.0	14.7	4.6	2.2	3.5	8.1	2.1	4.4	12.6
Montana	0.8	24.3	21.4	0.9	6.7	3.6	4.9	3.4	9.2
Nebraska	5.4	25.1	19.9	9.6	3.9	3.6	3.0	3.7	12.3
Nevada	11.2	1.2	8.2	5.0	10.3	12.6	1.6	4.3	11.6
New Hampshire	9.6	1.6	3.9	9.5	5.3	27.8	16.2	9.0	13.1
New Jersey	1.5	7.6	6.0	4.2	13.3	2.6	10.1	9.7	6.4
New Mexico	9.7	33.9	36.5	12.0	8.7	4.4	2.4	3.2	14.6
New York	21.2	4.8	5.5	1.6	15.1	9.1	8.6	3.1	4.8
North Carolina	10.6	6.6	4.3	1.1	7.3	4.5	6.6	2.5	9.7
North Dakota	3.9	24.8	14.5	3.4	3.3	2.6	1.1	2.8	3.6
Ohio	8.9	12.9	17.8	3.8	20.7	21.5	1.6	7.1	13.0
Oklahoma	6.7	9.9	10.3	3.3	0.8	4.8	9.8	5.0	10.2
Oregon	1.1	11.1	15.8	24.8	15.3	13.1	1.4	2.5	13.2
Pennsylvania	13.2	4.9	9.6	8.1	0.4	13.3	2.3	4.9	11.6
Rhode Island	18.7	0.4	1.9	2.3	10.0	20.4	22.8	8.5	26.4
South Carolina	7.3	6.4	8.8	3.5	3.6	3.0	5.4	3.9	2.8
South Dakota	4.0	17.9	11.8	2.2	3.3	0.4	2.1	6.0	5.7
Tennessee	10.5	2.3	11.4	16.3	7.2	12.2	4.0	2.9	9.6
Texas	6.7	2.6	5.0	0.8	7.8	0.8	5.1	4.6	0.8
Utah	2.3	4.4	15.4	22.0	22.9	1.3	7.9	2.1	19.7
Vermont	18.6	5.2	1.6	21.2	0.9	24.4	13.3	11.0	24.0
Virginia	11.9	6.1	0.9	2.2	0.5	3.5	1.1	0.8	11.3
Washington	1.8	3.1	11.1	12.6	13.7	4.6	11.1	27.5	19.9
West Virginia	5.3	11.7	10.3	1.0	9.0	11.2	2.4	21.8	27.8
Wisconsin	7.3	26.1	13.6	2.5	0.9	6.3	1.4	1.0	9.3
Wyoming	5.2	37.5	28.6	6.1	5.3	6.8	2.1	2.1	12.0

Supplementary Table 3 | MAPE (%) for 4-week forecasts of deaths for all the Japan prefectures on different prediction dates.

Forecast Date \Rightarrow	2020-11-15	2020-11-22	2020-11-30	2020-12-06	2020-12-13	2020-12-20	2020-12-27	2021-01-03	2021-01-10
Prefecture Name \Downarrow									
Aichi	14.2	9.2	12.6	19.3	25.8	27.2	11.0	1.4	3.9
Akita	26.1	67.2	46.2	19.0	24.0	13.2	117.2	115.3	37.4
Aomori	25.7	20.7	7.6	3.7	4.4	11.9	3.2	9.8	4.3
Chiba	2.1	1.2	11.0	12.9	4.1	2.8	12.0	14.7	8.2
Ehime	7.9	24.6	3.0	36.6	26.0	14.8	16.0	10.7	48.4
Fukui	6.2	13.7	9.8	4.3	5.9	7.1	9.4	31.6	29.9
Fukuoka	0.1	2.8	4.6	6.3	12.5	15.4	9.0	17.4	14.5
Fukushima	7.1	8.0	51.5	51.9	52.3	49.8	0.3	2.8	13.4
Gifu	9.9	13.1	31.7	39.7	40.9	45.2	21.3	33.5	39.2
Gunma	12.0	7.7	10.7	31.4	25.6	31.4	13.3	2.9	35.0
Hiroshima	3.4	19.3	17.8	66.6	37.1	38.6	5.2	27.2	16.0
Hokkaido	20.3	22.0	19.8	17.7	3.8	6.5	5.8	9.7	11.1
Hyogo	0.1	8.8	26.9	42.6	34.0	15.6	10.5	17.7	25.4
Ibaraki	9.6	17.4	1.1	13.6	8.2	0.5	0.3	15.0	6.8
Ishikawa	1.8	1.6	4.6	4.9	0.9	2.4	0.6	2.0	8.6
Iwate	54.4	61.9	70.3	74.5	39.5	6.7	13.9	0.9	8.8
Kagawa	8.0	11.8	4.2	40.5	5.4	30.7	48.1	65.9	45.7
Kagoshima	43.8	17.0	13.2	0.6	39.6	11.9	10.3	1.7	6.4
Kanagawa	0.6	3.8	11.1	11.1	6.3	6.0	2.9	17.2	8.7
Kochi	4.1	12.9	25.7	21.8	82.5	36.0	15.4	16.2	0.3
Kumamoto	18.9	37.2	7.7	8.9	39.4	52.0	35.6	32.3	39.4
Kyoto	8.4	17.6	12.1	12.1	26.0	24.0	7.0	20.5	15.8
Mie	13.2	3.5	5.1	18.2	24.5	20.9	17.3	14.0	49.0
Miyagi	57.8	63.5	3.7	13.6	20.1	13.5	14.3	23.2	14.4
Miyazaki	46.2	36.0	23.7	9.8	39.4	47.9	33.8	53.8	265.5
Nagano	117.7	76.5	20.6	44.8	34.5	31.1	25.1	29.2	39.7
Nagasaki	10.0	17.5	35.1	16.4	45.6	66.7	55.8	36.6	27.7
Nara	49.2	20.5	21.7	19.7	15.2	15.5	14.4	11.5	11.2
Niigata	–	88.5	13.3	1.5	211.6	23.4	3.4	17.9	40.4
Oita	10.2	43.1	41.8	39.3	15.2	39.3	46.4	42.8	12.1
Okayama	100.3	140.1	46.7	10.3	5.8	26.6	122.9	7.6	37.5
Okinawa	1.2	0.7	3.2	5.4	1.6	4.6	4.0	8.9	13.8
Osaka	19.1	16.5	0.7	19.0	18.9	24.4	7.2	4.2	25.6
Saga	31.8	26.9	22.9	11.7	44.7	17.9	28.3	20.9	38.0
Saitama	7.7	12.4	0.6	2.6	0.7	1.5	1.7	12.2	15.4
Shiga	23.1	21.5	1.9	4.7	23.3	32.1	20.4	10.5	1.1
Shizuoka	58.4	12.5	39.3	0.5	17.9	23.6	17.7	7.5	5.5
Tochigi	50.7	151.7	7.6	28.7	12.8	53.3	46.1	36.0	30.7
Tokushima	7.4	9.8	4.7	2.9	1.5	0.8	22.4	33.8	23.8
Tokyo	3.4	3.0	10.1	9.4	15.6	10.1	10.1	207.6	16.0
Tottori	–	–	–	–	13.4	63.5	43.9	19.6	102.4
Toyama	1.8	7.8	5.2	5.0	3.1	3.1	5.0	5.1	34.5
Wakayama	26.3	33.1	51.8	92.8	11.0	21.3	29.0	21.8	11.2
Yamagata	78.9	7.7	34.3	47.0	36.8	51.7	45.9	4.0	35.3
Yamaguchi	234.8	386.7	68.0	1.5	49.1	15.0	9.6	59.5	74.6
Yamanashi	3.6	4.3	3.3	4.1	3.7	0.9	7.1	9.9	11.1

Supplementary Table 4 | MAPE (%) for 4-week forecasts of confirmed cases for all the Japan prefectures on different prediction dates.

Forecast Date ==> Prefecture Name ↓	2020-11-15	2020-11-23	2020-11-30	2020-12-06	2020-12-13	2020-12-20	2020-12-27	2021-01-03	2021-01-10
Aichi	16.8	3.3	7.1	7.4	4.0	0.2	2.2	2.9	7.5
Akita	12.8	5.0	5.9	25.0	42.3	51.4	32.8	20.8	12.0
Aomori	7.2	11.4	1.5	2.1	31.8	38.5	22.4	3.4	7.4
Chiba	7.9	3.9	9.3	18.2	18.3	15.6	16.7	4.1	30.5
Ehime	48.1	44.3	21.6	4.2	42.3	53.0	43.2	36.2	10.1
Fukui	7.1	16.1	11.6	0.9	15.8	23.9	20.3	17.9	9.7
Fukuoka	8.6	10.1	13.5	10.1	11.5	12.9	3.1	30.2	0.9
Fukushima	6.6	5.7	26.0	35.0	42.8	45.4	19.8	15.5	2.4
Gifu	28.6	25.4	25.6	10.3	45.4	47.9	33.7	3.4	21.0
Gunma	29.9	22.1	13.9	7.6	35.0	38.6	21.3	23.9	3.7
Hiroshima	41.3	53.7	53.2	57.1	55.8	26.7	63.1	5.5	18.8
Hokkaido	30.7	87.5	51.2	11.1	24.2	3.5	7.4	1.8	8.7
Hyogo	24.9	5.6	47.6	7.5	5.8	1.8	9.9	12.3	12.2
Ibaraki	25.1	22.8	12.1	11.2	32.5	41.9	34.0	23.9	4.4
Ishikawa	3.6	1.1	3.9	8.8	23.6	25.7	11.9	5.1	0.1
Iwate	35.0	14.9	5.9	27.4	21.2	25.7	4.9	0.3	8.0
Kagawa	29.6	6.8	28.7	18.6	51.1	59.8	29.6	36.0	9.2
Kagoshima	6.5	14.6	22.0	29.4	24.5	31.1	25.7	23.9	10.8
Kanagawa	6.8	4.8	4.8	23.4	20.3	17.4	6.2	6.2	24.2
Kochi	54.3	64.8	70.5	20.9	46.6	32.0	12.4	5.9	1.2
Kumamoto	1.4	3.6	26.6	31.4	45.5	50.1	10.3	3.6	28.4
Kyoto	16.3	15.7	14.3	31.2	43.2	44.1	35.0	0.0	2.1
Mie	25.9	2.4	8.6	0.8	25.8	33.8	25.5	28.8	0.9
Miyagi	8.0	11.2	23.2	24.7	37.1	37.9	20.2	10.4	0.4
Miyazaki	28.2	5.3	13.9	10.6	49.4	53.9	46.1	51.1	47.2
Nagano	19.0	9.4	2.7	8.2	34.2	45.3	36.2	30.3	11.7
Nagasaki	8.0	17.1	21.0	55.7	64.6	68.8	58.1	30.7	98.8
Nara	19.3	9.8	16.3	10.0	33.0	34.4	16.6	6.5	0.3
Niigata	25.3	15.0	11.3	25.3	34.4	41.0	28.6	10.5	2.6
Oita	64.1	32.7	12.0	37.3	33.2	36.5	33.6	20.2	7.2
Okayama	5.3	41.7	8.1	36.4	55.0	42.5	24.1	18.0	22.1
Okinawa	2.9	1.6	4.5	7.7	16.2	20.8	15.7	13.6	4.8
Osaka	0.7	42.4	50.6	8.7	10.3	3.4	13.4	13.6	8.6
Saga	14.1	10.6	12.9	14.3	32.3	44.6	42.8	39.6	10.8
Saitama	9.3	5.1	10.3	3.5	1.9	2.3	5.8	1.1	18.1
Shiga	0.8	9.1	11.2	25.2	42.0	45.8	19.8	4.1	1.3
Shimane	6.5	9.8	14.2	7.4	24.7	21.7	26.6	0.5	4.1
Shizuoka	50.1	12.3	65.3	19.7	33.1	36.7	3.9	19.0	8.1
Tochigi	26.4	13.5	27.8	27.2	62.0	66.2	51.5	16.1	50.9
Tokushima	9.0	24.4	8.9	3.2	16.2	33.0	41.2	42.4	27.6
Tokyo	1.3	17.2	3.3	4.6	9.4	17.1	4.7	67.1	46.8
Tottori	0.1	2.0	14.0	12.2	58.6	60.2	27.1	11.1	9.1
Toyama	1.7	10.8	2.8	10.7	29.7	36.6	27.8	21.6	12.6
Wakayama	12.3	4.0	14.0	32.8	24.5	31.7	31.2	22.3	1.7
Yamagata	49.0	49.3	36.5	11.9	34.6	31.2	5.1	3.9	2.2
Yamaguchi	1.7	44.2	4.1	20.9	35.5	42.4	37.4	28.4	22.7
Yamanashi	19.8	16.4	16.0	14.3	34.2	42.6	33.4	18.8	1.1

Prospective comparisons to other models in the USA

We perform performance comparisons to the models submitted to "covid19-forecast-hub" (<https://covid19forecasthub.org/>). The repository accepts submissions for 4 targets (i) N-week ahead cumulative death, (ii) N-week ahead incremental death, (iii) N-week ahead incremental confirmed cases and (iv) N-day ahead incremental hospitalized. The submissions are up to 4 weeks of forecasting horizon. For each prediction date, we download the corresponding uploaded forecasts, and at the end of the horizon of 4 weeks, we perform the evaluations based on the ground truth on that day.

Tables 5, 6, 7, 10, 11 show the performance of our model and other model across different prediction dates. Extended Data Figures ?? and 2 and Figures 3, 5 and 6 show the corresponding normalized rankings. Overall, we observe that our model is consistently among the top models. Due to different numbers of available in each week, we consider the normalized rankings.

The comparison models are based on various epidemiological, statistical and machine learning approaches. The descriptions of the models are provided at "covid19-forecast-hub" (<https://covid19forecasthub.org/>), and summarized belows:

- **BPagano-RtDriven:** An SIR model for deaths and cases.
- **CEID-Walk:** A random walk model without drift, using the last few observations of a target time series to estimate the variance in step size of the random walk.
- **COVIDhub-baseline:** Extrapolation with the condition that the median prediction at all future horizons is equal to the most recent observed incidence.
- **COVIDhub-ensemble:** An ensemble, or model average, of submitted forecasts to the COVID-19 Forecast Hub.
- **Covid19Sim-Simulator:** An SEIR model with time-varying inputs, as well as state-specific estimates effective reproductive numbers.
- **CU-select:** A metapopulation SEIR model for projecting future COVID-19 incidence and deaths, using the scenario believed to be most plausible given the current setting.
- **Columbia_UNC-SurvCon:** A survival-convolution model with piece-wise transmission rates that incorporates latent incubation period and provides time-varying effective reproductive number.
- **CovidAnalytics-DELPHI:** An SEIR model augmented with underdetection and interventions, with the projections accounting for reopening and assuming interventions would be re-enacted if cases continue to climb.
- **DDS-NBDS:** Jointly modeling deaths and cases using a negative binomial distribution based nonparametric Bayesian generalized linear dynamical system.
- **Epiforecasts-ensemble1:** An ensemble of three different models based on forecasts for Rt, deaths only and a using deaths and cases.

-
- **GT-DeepCOVID:** A data-driven deep learning model learns the dependence of hospitalization and mortality rate on various detailed syndromic, demographic, mobility and clinical data (such as increase in positive Covid tests etc) from historical data.
 - **IHME-CurveFit:** Uses non-linear mixed effects curve-fitting with the assumption that current interventions continue.
 - **JCB-PRM:** The Political Realities Model (PRM) is built on observations of macro-level societal and political responses to COVID measured only in terms of infections and deaths.
 - **JHUAPL-Bucky:** A spatial compartment model using public mobility data with estimated local parameters (case reporting rates, doubling times, etc).
 - **JHU_UNC_GAS-StatMechPool:** This model jointly trains over locations, fitting to both mechanistic and statistical models, that enforces similar places having similar mechanistic coefficients.
 - **Karlen-pypm:** A population modeling framework using discrete-time difference equations with long periods of constant transmission rate.
 - **LANL-GrowthRate:** A statistical dynamical growth model accounting for population susceptibility, unconditional on particular intervention strategies.
 - **LNQ-ens1:** An ensemble of boosted tree and neural net models with engineered features.
 - **MITCovAlliance-SIR:** An SIR model trained on public health regions with parameters modeled as functions of static demographic and time-varying mobility features.
 - **MOBS-GLEAM_COVID:** A metapopulation, age structured compartmental model with the assumption that current interventions continue, modeling human mobility on a network.
 - **MSRA-DeepST:** A deep spatio-temporal network with knowledge based SEIR as a regularizer under the assumption of spatio-temporal process in pandemic of different regions.
 - **NotreDame-mobility:** An ensemble of nine models (each is a deterministic, SEIR-like model) that are identical except that they are driven by different mobility indices.
 - **OliverWyman-Navigator:** A compartmental model with non-stationary transition rates that incorporate: (1) the modeled effects of physical distancing and government actions; (2) evolving testing practices in terms of timeliness and scale; (3) the link to tracking and contact tracing, and (4) other region-specific characteristics.
 - **PSI-DRAFT:** A stochastic/deterministic, single-population SEIR model that stratifies by both age distribution and disease severity and includes generic intervention fitting.
 - **PI_UW-Mob_Collision:** A mobility-informed simplified SIR model motivated by collision theory.

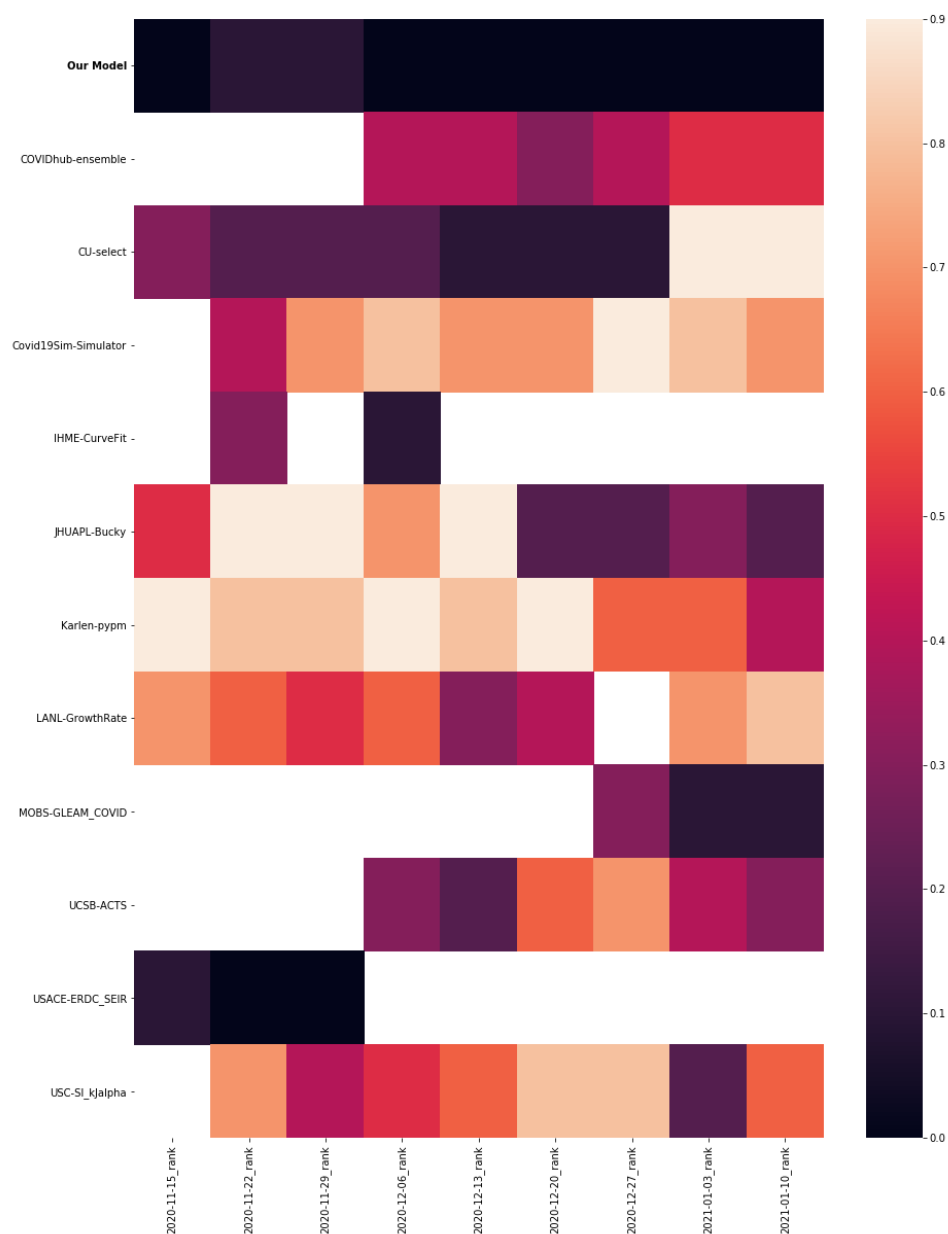
-
- **RobertWalraven-ESG**: A model based on fitting multiple skewed Gaussian distribution peaks.
 - **SteveMcConnell-CovidComplete**: A near-term fatality prediction model that uses estimated fatality trends, trends in positive virus tests and total virus tests, and age-related demographics.
 - **UA-EpiCovDA**: An SIR mechanistic model with parameters fitted using a variational data assimilation method.
 - **UCLA-SuEIR**: An SEIR model considering both untested and unreported cases, considering reopening and assumes susceptible population will increase after the reopen.
 - **UCSB-ACTS**: A data-driven machine learning model makes predictions by referring to other regions with similar growth patterns and assuming the similar development will take place in the current region.
 - **UCSD_NEU-DeepGLEAM**: A forecasting model that combines the signal of a discrete stochastic epidemic computational model with a deep learning spatiotemporal forecasting framework.
 - **UMass-MechBayes**: A Bayesian compartmental model with observations on cumulative case counts and deaths, fitted independently to each state, including observation noise and a case detection rate.
 - **UMich-RidgeTfReg**: A ridge regression (penalized Ordinary Least Squares regression) model to make predictions about future confirmed case and mortality trends.
 - **USC-SI_kJalpha**: A heterogeneous infection rate model with human mobility for epidemic modeling, adapting to changing trends and provide predictions of confirmed cases and deaths.
 - **UT-Mobility**: A Bayesian multilevel negative binomial regression model that uses the social distancing patterns from anonymized mobile-phone GPS traces.
 - **WalmartLabsML-LogForecasting**: A logistic growth prophet forecasting model fit using case counts and deaths as features.

Supplementary Table 5 | 4-week ahead MAE on different prediction dates for cumulative deaths for US states.

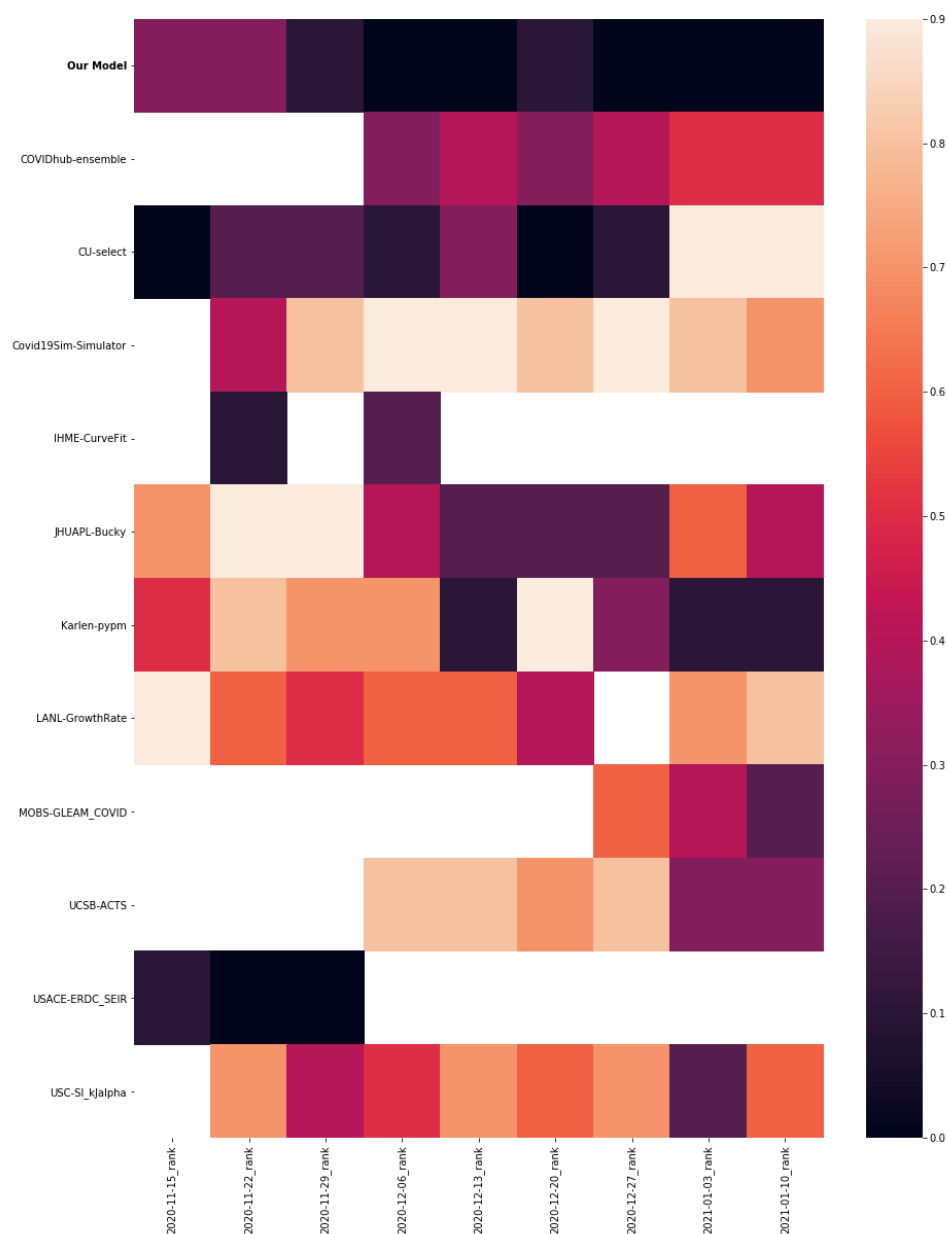
Model	11-15	11-22	11-29	12-06	12-13	12-20	12-27	01-03	01-10
Our model	242.2	189.3	218.8	242.6	251.6	215.1	306.5	359.9	332.6
BPagano-RtDriven	278.0	303.8	279.6	430.2	377.9	312.7	276.2	508.5	567.3
CEID-Walk	420.4	390.1	501.5	324.3	—	—	486.4	495.2	407.9
COVIDhub-baseline	420.3	386.8	503.9	322.7	379.1	374.4	496.2	513.7	392.2
COVIDhub-ensemble	268.8	195.4	324.6	225.7	239.5	210.1	359.0	453.4	321.3
CU-select	229.5	284.7	505.0	310.4	366.2	425.0	753.0	832.0	747.7
Columbia_UNC-SurvCon	697.3	536.5	392.2	—	—	—	422.0	730.0	—
Covid19Sim-Simulator	—	430.9	427.8	315.1	377.9	325.5	346.4	385.0	581.0
CovidAnalytics-DELPHI	481.9	420.1	399.4	393.5	435.3	292.5	577.7	654.2	852.5
DDS-NBDS	459.2	418.1	614.6	362.3	939.4	447.3	625.5	585.1	702.7
Epiforecasts-ensemble1	—	—	516.1	242.2	269.0	249.2	434.5	—	371.0
GT-DeepCOVID	—	—	—	229.5	313.2	228.7	385.7	—	323.2
IHME-CurveFit	—	312.8	—	338.0	—	—	—	—	—
JCB-PRM	240.9	261.9	280.2	313.1	337.3	339.7	277.7	—	—
JHUAPL-Bucky	351.4	423.2	325.4	264.6	302.4	543.3	780.8	687.1	743.3
JHU_CSSE-DECOM	242.2	257.4	291.3	352.8	404.0	410.0	394.7	—	341.3
Karlen-pypm	257.1	242.3	242.4	256.5	253.5	188.7	269.9	347.2	293.5
LANL-GrowthRate	308.2	322.4	471.1	304.4	286.6	446.9	—	533.2	457.8
LNQ-ens1	198.7	158.0	432.1	231.4	234.2	211.0	582.5	466.9	334.2
MIT_CritData-GBCF	—	—	—	310.1	333.7	378.7	400.6	522.1	400.9
MOBS-GLEAM_COVID	308.0	281.5	352.8	361.4	457.3	267.1	322.2	394.1	339.2
MSRA-DeepST	340.4	210.6	312.2	284.4	—	—	—	—	—
NotreDame-mobility	517.8	—	555.2	622.4	848.9	—	—	—	—
OliverWyman-Navigator	256.0	228.6	290.8	310.9	369.9	214.2	356.2	314.0	461.6
PSI-DRAFT	524.1	576.7	613.5	590.7	938.7	621.4	584.9	583.4	760.7
RobertWalraven-ESG	419.5	438.7	727.5	628.7	518.5	372.5	742.3	698.1	389.0
SteveMcConnell-CovidComplete	248.5	271.5	304.7	289.8	229.3	225.3	520.6	412.4	389.1
UA-EpiCovDA	292.0	298.8	492.2	298.6	272.7	482.2	900.8	562.1	307.3
UCLA-SuEIR	561.2	359.8	558.1	326.0	412.7	405.6	696.2	575.0	423.1
UCSB-ACTS	436.0	426.2	339.3	425.6	419.0	426.5	341.6	489.2	603.2
UCSD_NEU-DeepGLEAM	294.7	273.5	349.1	387.5	485.1	286.5	380.4	396.1	358.9
UMass-MechBayes	237.1	207.0	363.0	203.1	211.2	193.9	538.9	427.0	312.8
UMich-RidgeTfReg	443.5	471.6	—	—	—	544.3	625.4	542.9	724.2
USACE-ERDC_SEIR	1079.8	1278.2	1350.1	—	—	—	—	—	—
USC-SI_kJalpha	233.0	215.5	258.3	246.2	207.3	208.2	328.2	546.9	399.9
UT-Mobility	416.3	440.6	486.4	473.5	371.7	392.8	—	—	452.9
WalmartLabsML-LogForecasting	758.0	898.3	1001.2	1022.3	1087.2	1177.7	—	1249.9	1201.1



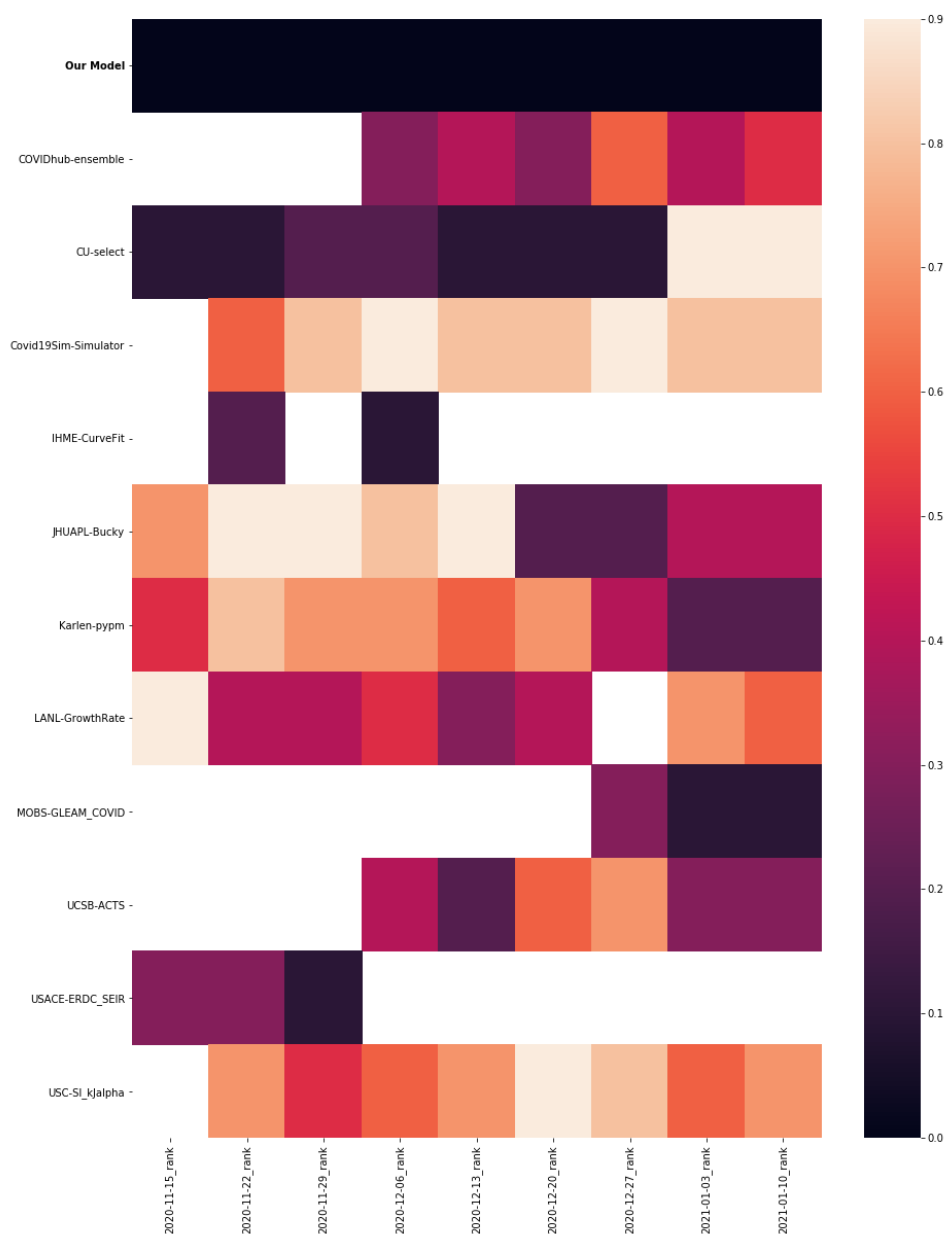
Supplementary Figure 2 | Normalized ranking for death WIS. The darker the color, the higher the ranking of the model is for the corresponding prediction date.



Supplementary Figure 3 | Normalized ranking for hospitalized increase MAE. The darker the color, the higher the ranking of the model is for the corresponding prediction date.



Supplementary Figure 4 | Normalized ranking for hospitalized increase MAPE. The darker the color, the higher the ranking of the model is for the corresponding prediction date.



Supplementary Figure 5 | Normalized ranking for hospitalized increase WIS. The darker the color, the higher the ranking of the model is for the corresponding prediction date.



Supplementary Figure 6 | Normalized ranking for accumulative cases MAE. The darker the color, the higher the ranking of the model is for the corresponding prediction date.



Supplementary Figure 7 | Normalized ranking for accumulative cases MAPE. The darker the color, the higher the ranking of the model is for the corresponding prediction date.

Supplementary Table 6 | 4-week ahead MAPE on different prediction dates for cumulative deaths for US states.

Model	11-15	11-22	11-29	12-06	12-13	12-20	12-27	01-03	01-10
Our model	5.4	6.0	5.6	5.1	5.5	4.5	4.5	4.9	5.9
BPagano-RtDriven	8.1	9.2	7.4	11.4	10.6	7.2	5.0	5.9	7.1
CEID-Walk	10.4	8.7	9.4	5.3	–	–	6.2	6.8	5.6
COVIDhub-baseline	10.5	8.8	9.3	5.6	7.0	5.3	5.9	6.8	5.8
COVIDhub-ensemble	7.0	5.3	6.0	3.9	3.9	3.8	4.3	5.2	4.5
CU-select	7.0	7.1	11.0	6.8	7.0	7.0	10.3	11.3	9.6
Columbia_UNC-SurvCon	19.0	14.4	9.4	–	–	–	7.0	8.9	–
Covid19Sim-Simulator	–	10.8	11.7	6.7	6.4	5.5	6.4	5.8	6.9
CovidAnalytics-DELPHI	11.3	7.0	6.6	6.4	6.2	4.7	7.2	7.5	9.1
DDS-NBDS	11.8	9.1	11.5	6.1	17.5	6.1	9.3	7.2	7.9
Epiforecasts-ensemble1	–	–	8.9	6.7	5.2	4.9	5.9	–	5.9
GT-DeepCOVID	–	–	–	4.8	6.4	4.3	5.4	–	4.8
IHME-CurveFit	–	8.5	–	9.8	–	–	–	–	–
JCB-PRM	7.2	7.4	6.7	5.8	5.8	6.6	5.2	–	–
JHUAPL-Bucky	8.3	10.6	8.4	5.1	6.2	8.1	10.8	7.7	8.3
JHU_CSSE-DECOM	6.1	8.6	6.7	6.9	9.6	8.2	4.1	–	5.5
Karlen-pypm	8.6	7.0	5.8	5.2	6.0	5.2	5.2	5.7	5.7
LANL-GrowthRate	9.9	8.3	9.7	7.0	5.8	5.8	–	6.9	6.3
LNQ-ens1	7.3	5.9	9.4	5.1	5.3	4.4	8.6	6.2	4.8
MIT_CritData-GBCF	–	–	–	7.2	8.3	9.5	5.4	7.5	6.7
MOBS-GLEAM_COVID	8.2	7.5	6.7	7.3	7.9	4.4	4.4	5.3	5.9
MSRA-DeepST	7.9	5.5	6.4	5.4	–	–	–	–	–
NotreDame-mobility	15.3	–	12.8	13.9	12.4	–	–	–	–
OliverWyman-Navigator	6.9	6.4	6.8	4.8	6.8	3.8	5.3	4.8	6.6
PSI-DRAFT	12.7	12.0	10.6	9.3	15.3	8.7	7.7	6.9	7.4
RobertWalraven-ESG	9.9	12.3	18.2	13.0	11.2	6.5	9.5	8.6	5.8
SteveMcConnell-CovidComplete	7.8	7.8	7.0	5.4	3.9	3.6	6.9	5.4	5.0
UA-EpiCovDA	7.7	6.6	9.1	6.9	5.1	6.6	11.0	6.8	5.7
UCLA-SuEIR	14.8	8.5	11.3	6.6	6.9	5.9	8.8	7.3	6.0
UCSB-ACTS	10.1	11.8	9.0	8.3	7.9	8.6	5.1	6.7	9.3
UCSD_NEU-DeepGLEAM	7.5	7.1	6.2	8.9	10.4	6.6	8.2	6.5	7.5
UMass-MechBayes	7.1	6.2	7.5	4.7	4.3	3.1	6.9	5.2	3.9
UMich-RidgeTfReg	11.7	11.8	–	–	–	7.6	7.6	7.4	7.9
USACE-ERDC_SEIR	18.5	19.8	19.5	–	–	–	–	–	–
USC-SI_kJalpha	7.6	6.6	6.7	4.9	4.8	3.6	5.5	7.2	5.6
UT-Mobility	12.4	12.3	11.8	11.3	8.5	7.7	–	–	5.8
WalmartLabsML-LogForecasting	19.1	20.0	19.7	17.7	16.0	15.4	–	13.6	12.2

Supplementary Table 7 | 4-week ahead WIS on different prediction dates for cumulative deaths for US states.

Model	11-15	11-22	11-29	12-06	12-13	12-20	12-27	01-03	01-10
Our model	188.3	119.4	137.4	156.7	151.4	149.2	221.8	287.5	246.1
BPagano-RtDriven	206.8	210.9	198.7	304.8	266.8	218.5	200.1	332.1	373.7
CEID-Walk	339.5	293.3	377.8	234.4	–	–	357.2	366.7	263.1
COVIDhub-baseline	304.9	270.0	363.8	223.4	280.5	263.0	381.5	398.5	269.7
COVIDhub-ensemble	180.9	130.8	214.8	163.9	156.8	132.5	236.0	329.6	210.9
CU-select	150.1	195.7	388.9	208.8	248.2	295.4	655.5	714.7	642.8
Columbia_UNC-SurvCon	580.5	438.3	331.4	–	–	–	345.2	667.2	–
Covid19Sim-Simulator	–	393.4	384.1	288.3	349.0	298.8	317.5	360.1	555.3
CovidAnalytics-DELPHI	433.2	366.3	334.7	327.3	359.0	229.7	473.7	551.1	742.2
DDS-NBDS	360.1	311.1	500.2	275.6	483.3	288.7	384.3	445.0	349.3
Epiforecasts-ensemble1	–	–	358.9	189.3	5258.5	172.9	280.6	–	262.7
GT-DeepCOVID	–	–	–	178.0	261.0	173.4	332.9	–	272.7
IHME-CurveFit	–	212.4	–	216.0	–	–	–	–	–
JCB-PRM	177.6	186.6	197.2	208.2	203.5	218.8	183.3	–	–
JHUAPL-Bucky	249.5	298.5	236.6	190.4	214.8	352.7	521.7	614.7	637.1
JHU_CSSE-DECOM	190.2	191.8	219.6	262.2	285.7	302.9	306.2	–	241.9
Karlen-pypm	165.7	170.0	162.3	160.8	176.9	136.6	181.3	221.9	199.7
LANL-GrowthRate	201.5	208.7	327.6	194.6	191.4	281.5	–	377.0	319.2
LNQ-ens1	138.4	139.8	242.0	174.1	183.0	171.0	297.3	270.5	252.2
MIT_CritData-GBCF	–	–	–	368.6	352.6	367.9	579.2	712.1	513.5
MOBS-GLEAM_COVID	219.5	175.2	253.8	254.8	296.7	168.7	243.2	307.0	254.3
MSRA-DeepST	323.3	201.4	294.7	219.5	–	–	–	–	–
NotreDame-mobility	497.3	–	534.3	594.8	823.7	–	–	–	–
OliverWyman-Navigator	193.8	161.2	214.8	238.6	269.3	147.1	222.9	207.9	297.9
PSI-DRAFT	474.9	521.0	570.0	531.0	812.5	534.2	484.4	510.4	663.7
RobertWalraven-ESG	356.7	380.6	516.6	585.1	410.1	361.7	534.6	505.4	480.9
SteveMcConnell-CovidComplete	159.7	197.5	199.6	230.0	227.3	229.8	268.1	253.0	297.2
UA-EpiCovDA	236.8	207.8	298.1	240.8	206.2	417.3	820.2	535.8	278.4
UCLA-SuEIR	539.4	318.1	515.8	287.7	373.8	364.9	649.2	523.0	370.6
UCSB-ACTS	301.5	300.9	238.4	298.4	278.3	285.4	235.1	329.2	376.1
UCSD_NEU-DeepGLEAM	209.5	194.3	264.8	288.0	378.7	206.7	262.7	292.5	252.3
UMass-MechBayes	169.1	159.4	255.9	158.9	161.0	143.0	359.7	279.8	216.3
UMich-RidgeTfReg	436.0	463.2	–	–	–	534.2	615.3	532.4	711.8
USACE-ERDC_SEIR	1045.2	1238.9	1312.7	–	–	–	–	–	–
USC-SI_kJalpha	151.2	138.4	179.1	159.1	143.9	151.8	213.2	391.6	267.3
UT-Mobility	310.5	294.5	345.5	317.4	249.8	270.5	–	–	290.2
WalmartLabsML-LogForecasting	667.5	797.9	874.1	876.3	916.5	989.1	–	1064.9	1018.3

Supplementary Table 8 | 4-week ahead MAE on different prediction dates for cases for US states.

Model	11-15	11-22	11-29	12-06	12-13	12-20	12-27	01-03	01-10
Our Model	32053.4	42754.7	40504.0	21474.3	38310.2	45864.2	28127.9	22410.9	57046.0
BPagano-RtDriven	46502.7	47753.3	35789.3	37879.8	33760.0	36767.5	37741.4	25705.1	40548.5
CEID-Walk	30310.4	32858.6	35872.2	29242.3	—	—	22179.2	16194.0	45916.9
COVIDhub-baseline	30186.1	32688.6	35194.0	29530.5	27115.7	19689.5	21590.9	15804.5	45831.3
COVIDhub-ensemble	29291.9	29328.1	29999.2	23426.3	18883.6	19599.5	22122.7	13059.5	45476.5
CU-select	22185.1	24381.8	33547.4	28237.0	20688.7	18125.6	30703.1	85884.9	134493.4
Covid19Sim-Simulator	—	32174.8	35998.7	32512.8	23420.4	22177.8	23391.2	17421.4	44462.0
CovidAnalytics-DELPHI	82754.5	71214.7	64714.4	59964.2	44217.4	40829.7	36745.5	33436.0	86958.2
DDS-NBDS	113852.5	84996.1	36806.9	62098.0	36791.1	41596.5	599623.1	62666.2	155824.3
JCB-PRM	28825.6	44413.1	49676.0	43186.5	39003.0	50171.3	39588.7	—	—
JHUAPL-Bucky	76184.9	70077.5	31703.4	30754.8	49608.7	51071.9	52437.5	21992.8	50275.4
JHU_CSSE-DECOM	33493.1	47062.1	34593.4	35246.7	24389.1	19124.5	20825.6	—	—
JHU_UNC_GAS-StatMechPool	39695.8	—	—	34528.7	33894.0	25865.0	38167.2	19100.3	25647.2
Karlen-pypm	60896.6	65833.9	40988.6	37408.5	55336.1	53411.9	73077.0	25990.3	39715.5
LANL-GrowthRate	42174.5	28576.7	27922.2	17279.5	26617.2	34812.7	—	16757.3	48362.5
LNQ-ens1	29175.9	23637.3	31136.6	20736.0	19698.9	21867.1	33130.9	14730.3	59385.2
MOBS-GLEAM_COVID	—	—	—	—	—	—	—	—	18681.7
MSRA-DeepST	30740.6	38163.4	31998.4	32767.3	—	—	—	—	—
RobertWalraven-ESG	36836.8	32084.2	43150.7	36997.7	21797.6	23309.9	38841.3	18920.6	42033.6
TTU-squider	57551.1	62060.6	83003.0	50172.8	37273.9	30865.3	32628.1	35749.1	35148.6
UCLA-SuEIR	47681.5	35547.5	40870.7	36870.3	31336.7	24932.6	28068.2	23891.6	42193.1
UCSB-ACTS	33140.8	63182.9	88967.8	69733.6	57322.9	38127.6	71747.6	59871.0	49562.5
UMich-RidgeTfReg	89348.0	52521.7	—	—	—	59061.2	32540.8	35190.5	64749.4
USACE-ERDC_SEIR	22132.4	36251.4	53714.1	—	—	—	—	—	—
USC-SI_kJalpha	41908.5	32088.8	34601.3	22082.6	29471.4	24764.7	27827.5	19520.7	37291.5

Supplementary Table 9 | 4-week ahead MAPE on different prediction dates for cases for US states.

Model	11-15	11-22	11-29	12-06	12-13	12-20	12-27	01-03	01-10
Our Model	8.8	12.4	11.3	6.6	7.6	8.8	6.2	5.6	13.4
BPagano-RtDriven	15.5	21.4	11.2	13.8	10.4	7.0	6.3	4.8	8.0
CEID-Walk	9.7	9.1	8.9	7.2	–	–	4.0	3.1	8.3
COVIDhub-baseline	9.7	8.9	8.7	7.2	6.2	4.0	3.9	3.1	8.4
COVIDhub-ensemble	9.1	9.3	7.7	6.7	5.4	3.0	3.7	2.4	7.9
CU-select	8.1	7.7	7.9	7.5	5.1	3.6	7.6	17.2	21.7
Covid19Sim-Simulator	–	9.4	9.6	8.2	6.6	4.1	3.3	2.5	7.3
CovidAnalytics-DELPHI	28.1	30.3	21.8	17.7	12.4	6.4	7.6	7.0	13.7
DDS-NBDS	97.9	30.1	12.2	24.3	11.9	14.8	210.9	13.0	29.0
JCB-PRM	10.7	14.2	15.4	13.1	11.5	10.1	6.2	–	–
JHUAPL-Bucky	35.9	31.5	10.9	8.9	10.5	7.7	8.5	4.8	11.1
JHU_CSSE-DECOM	13.1	17.2	9.7	8.7	7.1	4.4	4.3	–	–
JHU_UNC_GAS-StatMechPool	12.9	–	–	8.8	7.8	4.7	8.1	4.4	5.4
Karlen-pypm	27.7	28.8	13.0	13.6	14.6	9.4	9.1	5.9	6.9
LANL-GrowthRate	13.9	10.7	7.1	5.1	5.4	4.3	–	3.6	8.6
LNQ-ens1	10.1	7.7	7.2	6.5	4.8	3.5	6.9	3.2	10.6
MOBS-GLEAM_COVID	–	–	–	–	–	–	–	–	5.1
MSRA-DeepST	10.4	13.6	8.1	8.2	–	–	–	–	–
RobertWalraven-ESG	13.9	10.6	13.3	9.1	6.5	4.4	8.1	4.1	7.3
TTU-squider	19.4	18.8	21.6	11.8	12.0	9.4	8.2	6.7	6.4
UCLA-SuEIR	15.0	9.4	9.7	9.7	7.5	5.5	5.5	4.6	7.1
UCSB-ACTS	11.6	20.1	28.4	22.7	17.6	10.2	12.7	9.6	8.1
UMich-RidgeTfReg	33.9	20.7	–	–	–	8.3	5.6	5.4	10.2
USACE-ERDC_SEIR	11.0	10.7	13.9	–	–	–	–	–	–
USC-SI_kJalpha	15.0	11.3	8.7	6.6	7.2	3.9	4.1	3.7	6.3

Supplementary Table 10 | 4-week ahead MAE on different prediction dates for hospitalized increased for US states.

Model	11-15	11-22	11-29	12-06	12-13	12-20	12-27	01-03	01-10
Our Model	74.8	85.8	82.3	107.5	136.4	110.6	140.1	85.6	55.2
COVIDhub-ensemble	–	–	–	279.3	368.5	269.1	281.2	244.4	323.4
CU-select	105.9	101.9	94.8	145.7	309.3	227.8	209.9	566.0	671.5
Covid19Sim-Simulator	–	247.2	265.0	419.3	429.3	323.9	376.6	354.0	369.3
IHME-CurveFit	–	106.0	–	140.5	–	–	–	–	–
JHUAPL-Bucky	264.5	842.5	401.3	373.3	518.6	237.9	239.5	236.0	277.1
Karlen-pypm	313.2	481.7	352.0	425.4	492.4	379.9	324.3	264.0	302.3
LANL-GrowthRate	303.0	257.7	231.2	300.1	365.1	295.8	–	298.0	373.8
MOBS-GLEAM_COVID	–	–	–	–	–	–	257.8	190.8	235.8
UCSB-ACTS	–	–	–	274.2	332.8	301.4	342.3	237.5	283.5
USACE-ERDC_SEIR	97.8	84.4	62.5	–	–	–	–	–	–
USC-SI_kJalpha	–	322.6	218.0	292.3	413.0	346.3	362.9	231.9	356.1

Supplementary Table 11 | 4-week ahead MAPE on different prediction dates for hospitalized increased for US states.

Model	11-15	11-22	11-29	12-06	12-13	12-20	12-27	01-03	01-10
Our Model	131.0	185.0	189.6	104.5	117.2	124.1	65.4	70.7	71.3
COVIDhub-ensemble	–	–	–	224.0	254.3	186.1	135.4	173.8	286.3
CU-select	80.9	97.3	219.9	109.2	237.8	117.7	69.2	488.5	806.8
Covid19Sim-Simulator	–	284.8	645.3	390.9	385.9	253.3	243.3	255.5	346.3
IHME-CurveFit	–	79.1	–	147.1	–	–	–	–	–
JHUAPL-Bucky	325.7	944.3	738.9	248.2	215.6	169.5	100.1	200.4	282.6
Karlen-pypm	298.4	835.6	532.8	289.0	215.1	324.8	109.3	139.4	226.7
LANL-GrowthRate	334.1	374.0	477.2	264.3	267.9	189.8	–	234.8	440.8
MOBS-GLEAM_COVID	–	–	–	–	–	–	141.1	153.8	267.1
UCSB-ACTS	–	–	–	294.7	382.0	247.5	188.5	148.7	272.1
USACE-ERDC_SEIR	128.0	71.0	106.8	–	–	–	–	–	–
USC-SI_kJalpha	–	459.4	458.5	250.3	330.6	245.6	172.8	144.4	309.6

Supplementary Table 12 | 4-week ahead WIS on different prediction dates for hospitalized increased for US states.

Model	11-15	11-22	11-29	12-06	12-13	12-20	12-27	01-03	01-10
Our Model	55.6	54.3	51.6	66.6	107.4	73.4	115.9	56.3	35.3
COVIDhub-ensemble	–	–	–	223.2	307.7	213.9	238.4	196.2	262.5
CU-select	88.7	81.8	79.5	121.9	268.9	194.3	180.5	441.5	556.7
Covid19Sim-Simulator	–	234.8	248.6	408.3	419.2	314.0	366.7	346.3	359.1
IHME-CurveFit	–	82.5	–	110.4	–	–	–	–	–
JHUAPL-Bucky	214.7	667.4	340.0	313.4	445.2	197.1	205.6	196.2	235.6
Karlen-pypm	210.7	333.5	241.6	294.0	351.1	263.4	237.2	178.8	209.1
LANL-GrowthRate	231.7	195.8	175.5	229.3	284.8	219.4	–	216.2	278.5
MOBS-GLEAM_COVID	–	–	–	–	–	–	221.2	150.4	182.6
UCSB-ACTS	–	–	–	223.3	276.6	243.0	282.4	179.6	226.7
USACE-ERDC_SEIR	95.8	82.8	61.3	–	–	–	–	–	–
USC-SI_kJalpha	–	285.2	195.7	266.6	384.4	319.5	321.6	206.1	317.1

Statistical comparison of forecasting accuracy

We compare our models' US state forecasts to those of alternative models' by applying the two-sided Diebold-Mariano (DM) test with the small sample size correction [2] to the MAE or MAPE averaged across all the locations for the dates when for both of the models produced forecasts. The DM test was used to compare the forecasts, rather than other statistical tests, because it accounts for the correlation of forecast errors across time [3]. Only models with more than 4 comparisons were included in the results. The p-values from the tests were adjusted using the Holm–Bonferroni method [4] to account for the multiple comparisons for each error. KPSS tests [5] were run on the differences in the loss metrics and rejections of the null hypothesis of stationary about either a constant level or a trend at the 0.05 level are also noted. For each of these tests, a positive DM statistic suggests that our models' forecasts were more accurate (i.e. had a smaller loss) than the competitor's forecasts while a negative statistic suggests that the inverse is true.

Supplementary Table 13 | Diebold-Mariano test results for 4-week ahead predictions of cumulative deaths for US states using MAE as the loss. Each row represents an alternative forecasting model compared with our model. A positive DM statistic suggests that our models' forecasts were more accurate (i.e. had a smaller loss) than the competitor's forecasts, while a negative statistic suggests that the inverse is true.

Model	Comparisons	DM Statistic	p-value	Adj. p-value	KPSS Rejected
Karlen-pypm	8	0.155	0.881	1	
UMass-MechBayes	8	1.278	0.242	1	
LNQ-ens1	8	1.400	0.204	1	
USC-SI_kJalpha	8	1.403	0.203	1	
Covid19Sim-Simulator	7	1.580	0.165	1	level
Epiforecasts-ensemble1	5	1.727	0.159	1	level
CU-select	8	1.737	0.126	1	
COVIDhub-ensemble	8	1.849	0.107	1	
JHU_CSSE-DECOM	7	1.886	0.108	1	
OliverWyman-Navigator	8	1.970	0.089	0.984	level
JHUAPL-Bucky	8	2.286	0.056	0.763	
SteveMcConnell-CovidComplete	8	2.306	0.055	0.763	
MIT_CritData-GBCF	5	2.405	0.074	0.888	level
UA-EpiCovDA	8	2.654	0.033	0.492	
JCB-PRM	7	3.239	0.018	0.283	level
MOBS-GLEAM_COVID	8	3.523	0.01	0.174	level
BPagano-RtDriven	8	3.810	0.007	0.126	
LANL-GrowthRate	7	4.174	0.006	0.117	
Columbia_UNC-SurvCon	5	4.594	0.01	0.174	
WalmartLabsML-LogForecasting	7	5.723	0.001	0.026	level
UCSD_NEU-DeepGLEAM	8	5.963	0.001	0.013	level
UCLA-SuEIR	8	6.305	< 0.001	0.01	level
UCSB-ACTS	8	6.597	< 0.001	0.008	level
RobertWalraven-ESG	8	7.168	< 0.001	0.005	
CovidAnalytics-DELPHI	8	7.641	< 0.001	0.004	
COVIDhub-baseline	8	8.157	< 0.001	0.002	
UMich-RidgeTfReg	5	8.747	0.001	0.021	
UT-Mobility	6	8.918	< 0.001	0.008	
CEID-Walk	6	9.357	< 0.001	0.006	level
PSI-DRAFT	8	11.047	< 0.001	< 0.001	
DDS-NBDS	8	16.886	< 0.001	< 0.001	level, trend

Supplementary Table 14 | Diebold-Mariano test results for 4-week ahead predictions of cumulative deaths for US states using MAPE as the loss. Each row represents an alternative forecasting model compared with our model. A positive DM statistic suggests that our models' forecasts were more accurate (i.e. had a smaller loss) than the competitor's forecasts, while a negative statistic suggests that the inverse is true.

Model	Comparisons	DM Statistic	p-value	Adj. p-value	KPSS Rejected
COVIDhub-ensemble	8	-0.559	0.594	1	level
UMass-MechBayes	8	0.892	0.402	1	
SteveMcConnell-CovidComplete	8	1.259	0.249	1	
USC-SI_kAlpha	8	1.423	0.198	1	
Columbia_UNC-SurvCon	5	1.519	0.203	1	
Covid19Sim-Simulator	7	1.549	0.172	1	
MOBS-GLEAM_COVID	8	1.831	0.11	1	
COVIDhub-baseline	8	2.028	0.082	1	level
LANL-GrowthRate	7	2.030	0.089	1	level
LNQ-ens1	8	2.117	0.072	1	level
Epiforecasts-ensemble1	5	2.131	0.1	1	
UMich-RidgeTfReg	5	2.134	0.1	1	level
Karlen-pypm	8	2.208	0.063	0.964	level
CEID-Walk	6	2.217	0.077	1	level
CovidAnalytics-DELPHI	8	2.238	0.06	0.964	
UCLA-SuEIR	8	2.331	0.053	0.893	level
UCSB-ACTS	8	2.742	0.029	0.548	level
UT-Mobility	6	3.003	0.03	0.548	level
OliverWyman-Navigator	8	3.328	0.013	0.265	level
CU-select	8	3.437	0.011	0.239	
DDS-NBDS	8	3.638	0.008	0.191	level
MIT_CritData-GBCF	5	3.991	0.016	0.325	level, trend
JHUAPL-Bucky	8	4.126	0.004	0.106	
RobertWalraven-ESG	8	4.431	0.003	0.076	
UCSD_NEU-DeepGLEAM	8	4.684	0.002	0.058	
BPagano-RtDriven	8	5.740	0.001	0.02	
WalmartLabsML-LogForecasting	7	5.885	0.001	0.029	
UA-EpiCovDA	8	6.334	< 0.001	0.011	level
JHU_CSSE-DECOM	7	7.947	< 0.001	0.006	
PSI-DRAFT	8	9.407	< 0.001	0.001	
JCB-PRM	7	10.311	< 0.001	0.002	level

Tables 13 and 14 show the DM test results for the four week ahead forecasts of cumulative deaths in the US for MAE and MAPE respectively. When MAE is considered, none of the DM statistics are negative, in other words our model forecasts have a lower MAE than any of the alternative models. This difference is statistically significant, when adjusted p-value threshold of 0.05 is considered, in 12 of the 30 comparisons. When MAPE is compared the 'COVIDhub-ensemble' forecasts have a negative DM statistic, suggesting that it has a slightly lower MAPE, but this difference is not statistically significant with a p-value greater than 0.5, even before the p-value adjustment. The better relative results of our model when evaluated with MAE vs. MAPE may be due to the fact that the loss function we used to train the model (see Methods), as it is not normalized with the absolute quantities.

When other top models are considered, we observe that they are often either in 'black-box' (non-explainable) ensemble model form (such as 'COVIDhub-ensemble' and 'LNQ-ens1'), or complex variants of compartmental models, improved with more advanced variable fitting, appropriate

prior assumptions and sometimes additional features similar to our model (e.g. 'USC-SI_kJalpha', 'Karlen-pypm', 'UMass-MechBayes'). Our explainable AI-augmented compartmental model, with its competitive accuracy, constitutes an appreciable alternative.

Supplementary Table 15 | Diebold-Mariano test results for 4-week ahead predictions of cumulative confirmed cases for US states using MAE as the loss. Each row represents an alternative forecasting model compared with our model. A positive DM statistic suggests that our models' forecasts were more accurate (i.e. had a smaller loss) than the competitor's forecasts, while a negative statistic suggests that the inverse is true.

Model	Comparisons	DM Statistic	p-value	Adj. p-value	KPSS Rejected
LNQ-ens1	8	-4.609	0.002	0.054	level
COVIDhub-ensemble	8	-3.387	0.012	0.21	
COVIDhub-baseline	8	-3.199	0.015	0.256	
Covid19Sim-Simulator	7	-1.841	0.115	1	
LANL-GrowthRate	7	-1.613	0.158	1	level
CEID-Walk	6	-1.385	0.225	1	
USC-SI_kJalpha	8	-1.312	0.231	1	
RobertWalraven-ESG	8	-1.085	0.314	1	
JHU_CSSE-DECOM	7	-0.578	0.584	1	level
CU-select	8	-0.105	0.919	1	
UCLA-SuEIR	8	-0.058	0.955	1	
JHU_UNC_GAS-StatMechPool	6	0.329	0.755	1	
BPagano-RtDriven	8	0.792	0.455	1	level
UMich-RidgeTfReg	5	1.224	0.288	1	level
DDS-NBDS	8	1.248	0.252	1	level
CovidAnalytics-DELPHI	8	1.291	0.238	1	
TTU-squider	8	1.375	0.211	1	
JHUAPL-Bucky	8	2.021	0.083	1	
JCB-PRM	7	2.069	0.084	1	
UCSB-ACTS	8	3.440	0.011	0.206	
Karlen-pypm	8	4.436	0.003	0.063	level, trend

Tables 15 and 16 show the DM test results for the four week ahead forecasts of US state confirmed cases for MAE or MAPE, respectively. Compared to our models' performance on cumulative deaths (see the previous section) our models' forecasts are relatively less accurate for confirmed cases in comparison to alternative models. This is largely due to our choice of overemphasizing on predicting deaths more accurately when the multi-objective optimization trade-off is considered (see Methods). As the data quality of the confirmed case counts is lower than the death counts (due to its dependence on the testing rates and the reporting differences across locations), we choose to tune our model to fit better to the death counts than the confirmed case counts. We indeed empirically demonstrate that by increasing the confirmed case count coefficient, we can improve confirmed case forecasting accuracy, albeit a decrease in death forecasting accuracy.

In general, relatively large amounts of variability in performance is observed among the models across the prospective time range. In fact, none of the forecasts is better than our forecasts in a statistically significant way for a significance level of 0.1 (the minimum p-value is 0.054, for 'LNQ-ens1') for MAE, and only only two models ('Covid19Sim-Simulator' and 'COVIDhub-baseline') are better than ours for MAPE, in a statistically significant way. Similar to cumulative death benchmarking, we observe that the two 'black-box' ensemble models,

Supplementary Table 16 | Diebold-Mariano test results for 4-week ahead predictions of cumulative confirmed cases for US states using MAPE as the loss. Each row represents an alternative forecasting model compared with our model. A positive DM statistic suggests that our models' forecasts were more accurate (i.e. had a smaller loss) than the competitor's forecasts, while a negative statistic suggests that the inverse is true.

Model	Comparisons	DM Statistic	p-value	Adj. p-value	KPSS Rejected
Covid19Sim-Simulator	7	-6.388	0.001	0.015	
COVIDhub-baseline	8	-2.967	0.021	0.397	level
COVIDhub-ensemble	8	-2.626	0.034	0.614	level
LNQ-ens1	8	-2.424	0.046	0.779	level
CEID-Walk	6	-1.709	0.148	1	level
LANL-GrowthRate	7	-0.763	0.474	1	
USC-SI_kJalpha	8	-0.435	0.677	1	
CU-select	8	-0.122	0.906	1	
UCLA-SuEIR	8	-0.040	0.97	1	level
JHU_CSSE-DECOM	7	0.136	0.897	1	level
RobertWalraven-ESG	8	0.224	0.829	1	level
JHU_UNC_GAS-StatMechPool	6	0.450	0.672	1	
UMich-RidgeTfReg	5	0.577	0.595	1	
JHUAPL-Bucky	8	0.833	0.432	1	
BPagano-RtDriven	8	0.881	0.407	1	level, trend
CovidAnalytics-DELPHI	8	1.046	0.33	1	
Karlen-pypm	8	1.240	0.255	1	
TTU-squider	8	1.468	0.186	1	level, trend
DDS-NBDS	8	2.061	0.078	1	level
JCB-PRM	7	3.446	0.014	0.274	
UCSB-ACTS	8	5.657	0.001	0.016	

'COVIDhub-ensemble' and 'LNQ-ens1', are among the top ones. All in all, such ensemble models constitute great alternatives when only the goal of forecasting accuracy is considered, yet their lack of explainability is a significant concern for epidemiological use. Among the other top models, 'Covid19Sim-Simulator' is an SEIR-variant that uses time-varying features, similar to ours. We note that 'Covid19Sim-Simulator' is also one of the top models for the cumulative death forecasts, supporting the importance of data-driven machine-learned compartmental modeling. Interestingly, 'COVIDhub-baseline', a simple trend following model, that does not have any complex machine learning components or that does not make use of any data, is one of the top models for cumulative case forecasting, suggesting that in the prospective evaluation period, the confirmed case increasing trends have similar patterns between consecutive weeks, and additional predictability is small. For death forecasting, 'COVIDhub-baseline' is one of the poorest models though. Vice versa, one of the top models for death forecasting, 'Karlen-pypm', is one of the worst for confirmed case forecasting.

Tables 17 and 18 show the DM tests between our model and alternative models, for 4-week ahead forecasts of the increase in hospitalized cases in US states, in MAE and MAPE respectively. This prediction is particularly important as it can be used for to plan and allocate hospital resources, and to prevent the overwhelming of the hospital systems. Similar to confirmed cases, this data is also observed to be lower quality than the data for the death counts.

Our models' errors are smaller than any of the alternative models using either MAE or MAPE and this difference is statistically significant for most models (when adjusted p-value<0.05

Supplementary Table 17 | Diebold-Mariano test results for 4-week ahead predictions of the increase in hospitalized cases for US states using MAE as the loss. Each row represents an alternative forecasting model compared with our model. A positive DM statistic suggests that our models' forecasts were more accurate (i.e. had a smaller loss) than the competitor's forecasts, while a negative statistic suggests that the inverse is true.

Model	Comparisons	DM Statistic	p-value	Adj. p-value	KPSS Rejected
CU-select	8	1.046	0.33	0.65	level
JHUAPL-Bucky	8	2.039	0.081	0.242	level
Karlen-pypm	8	5.892	0.001	0.002	
USC-SI_kJalpha	7	10.512	< 0.001	< 0.001	
COVIDhub-ensemble	5	10.992	< 0.001	0.002	level
Covid19Sim-Simulator	7	11.214	< 0.001	< 0.001	
LANL-GrowthRate	7	17.394	< 0.001	< 0.001	
UCSB-ACTS	5	18.954	< 0.001	< 0.001	level

Supplementary Table 18 | Diebold-Mariano test results for 4-week ahead predictions of the increase in hospitalized cases for US states using MAPE as the loss. Each row represents an alternative forecasting model compared with our model. A positive DM statistic suggests that our models' forecasts were more accurate (i.e. had a smaller loss) than the competitor's forecasts, while a negative statistic suggests that the inverse is true.

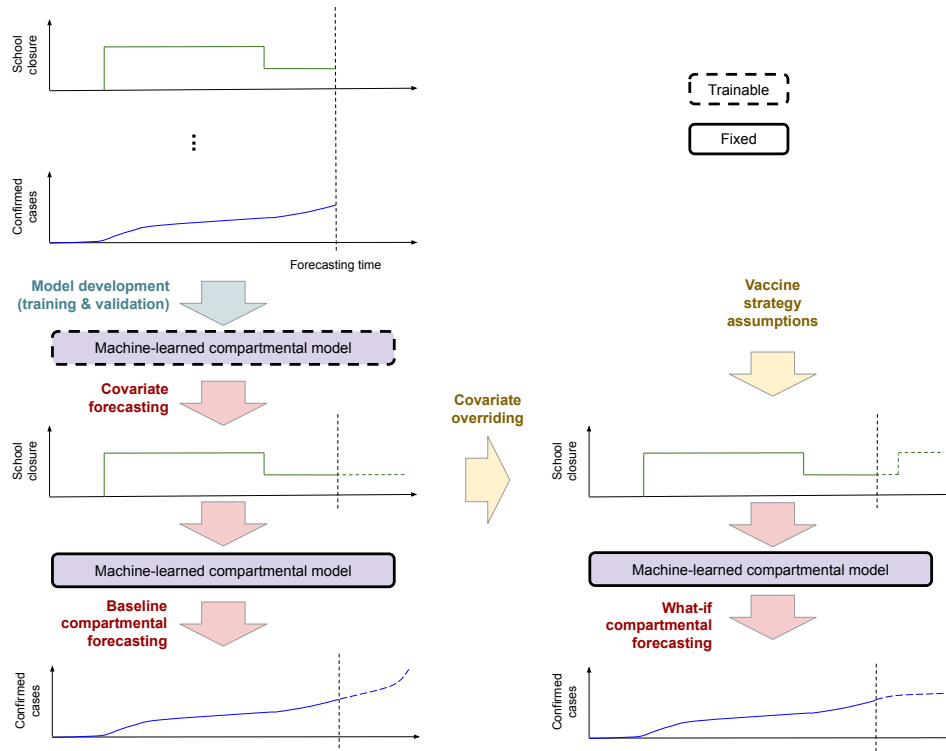
Model	Comparisons	DM Statistic	p-value	Adj. p-value	KPSS Rejected
CU-select	8	0.556	0.595	0.595	level
JHUAPL-Bucky	8	1.358	0.216	0.553	
Karlen-pypm	8	1.738	0.126	0.503	level
USC-SI_kJalpha	7	2.251	0.065	0.327	level
UCSB-ACTS	5	3.017	0.039	0.236	level
Covid19Sim-Simulator	7	5.016	0.002	0.020	level
LANL-GrowthRate	7	5.100	0.002	0.020	level
COVIDhub-ensemble	5	6.873	0.002	0.020	level

threshold is considered, for 75% of the models using MAE and 38% of the models using MAPE). 'CU-select' and 'JHUAPL-Bucky', SEIR-variant models that also utilize interventions are among the other top ones for the hospitalization predictions.

3 Counterfactual analysis

Figure 8 overviews the training and prediction with the proposed framework for counterfactual analysis. On each prediction date, using observed values for the case counts and features, a forecasting model is fitted. This forecasting model learns how to relate particular features to the compartmental outcomes, such as how a particular NPI would affect the number of exposed people. The forecasting model also needs the future values of the features, as it generates the predictions for all the time steps in the forecasting horizon conditioned on the values of the features in the forecasting horizon. For this purpose, we forecast the future values of the features, as explained in Methods. Using these forecasted values, we obtain 'baseline compartmental forecasts', in

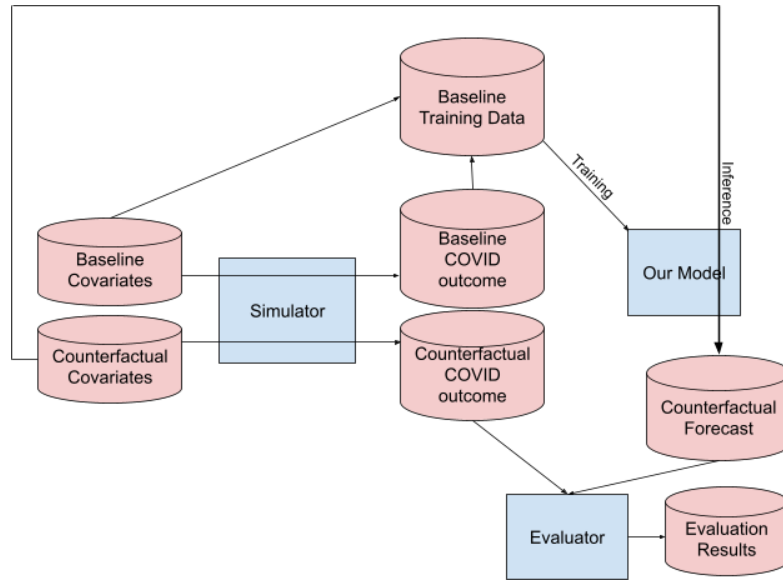
other words, the most likely outcomes given the past observations. For counterfactual analysis, the forecasted features are overridden. Essentially, we are trying to answer the question of what would happen if a particular feature changes in this particular way during the forecasting horizon? The overriding may be employed for any of the forecasted features.



Supplementary Figure 8 | Overview of model development, baseline forecasting and what-if forecasting for counterfactual analysis. Using the past-observed values, the proposed machine-learned compartmental model is fitted. For forecasting, the future values of features are needed with the proposed compartmental model, for which they are separately forecasted. These feature forecasts can be used for “baseline compartmental forecasts”, in other words, expected scenarios based on the current observations. Alternatively, they can be overriding the counterfactual assumptions on them, including the vaccine strategy related assumptions that are assumed to be non-existent for the baseline case. Eventually, with these overridden features, counterfactual scenarios for compartmental forecasts can be obtained.

Evaluation of Counterfactual Analysis

Evaluations of counterfactual scenarios are difficult, as we do not know the underlying ground truth associated with counterfactual scenarios. To demonstrate the reliability of our framework when it is used to analyze the impacts of NPIs, we rely on two evaluation methods. Note that the impact of vaccinations are determined by the way we modify the compartmental modeling, rather than counterfactual analysis.



Supplementary Figure 9 | Block diagram of counterfactual evaluation using simulated data. First, a forecasting model is trained to be used as the simulator. With baseline features and counterfactual features as inputs to this simulator, baseline and counterfactual simulated data are generated by the simulator. Baseline simulated data is used to train the forecasting model. Using counterfactual features, forecasting model generates counterfactual forecasts, and these forecasts are evaluated against the simulated counterfactual ground truth.

Using simulated data: Since outcomes of counterfactual interventions cannot be observed in the real world, our first evaluation approach is based on a simulator, that is capable of simulating outcomes for both factual and counterfactual scenarios. Our simulator uses features (such as NPIs) as inputs, and outputs confirmed cases and deaths. Using this simulator, data for both training and evaluation can be synthesized, and the forecasting model can be trained on the simulated training data. Then, the trained forecasting model can be used to generate counterfactual forecasts, and since we have access to the simulator that can also synthesize such counterfactual scenarios, we are able to directly quantify the counterfactual scenario generation reliability by comparing it to the simulated counterfactual outcomes. Figure 9 overviews the entire process of counterfactual evaluation using simulated data.

Evaluations on simulated data would be helpful in detecting either model problems (e.g., poor estimation of model weights) or data problems (e.g., school closures at a certain stage of COVID-19 did not happen sufficient number of times for the model to accurately learn their effect). As the simulator we use the proposed compartmental model, with fitted encoder weights and initial conditions. While fitting to observed data, we disable teacher forcing (see Methods) so that data-driven conditioning can be learned because ground truth data is not available for counterfactual inference, so there is nothing to condition on. To validate that the model without teacher forcing still fits the observed ground truth data reasonably well, we compare the original ground truth confirmed cases and deaths with the simulated data and found that MAPE was less

than 30% for all forecasting dates.

Ideally, the error of our counterfactual estimates should not be worse than the error of baseline factual estimates. To capture this, we propose the evaluation metric of the ratio between the 'factual MAPE' (MAPE between factual prediction and factual ground truth generated by the simulator) and the 'counterfactual MAPE' (MAPE between counterfactual prediction and counterfactual ground truth generated by the simulator). We evaluate this on 6 different counterfactual scenarios. Table 19 shows that for most scenarios and forecasting dates, the differences between factual and counterfactual MAPEs are less than 5%.

Supplementary Table 19 | Ratios between factual MAPE and counterfactual MAPE in 6 different scenarios for the US model.

Forecasting date	9/7		9/28		10/19	
Compartments	Confirmed	Deaths	Confirmed	Deaths	Confirmed	Deaths
Mobility Increase (x10)	1.014	1.011	1.002	1.000	1.091	0.998
Mobility Decrease (90%)	0.998	0.999	1.000	1.000	0.997	1.000
School Closure	1.017	1.018	0.999	1.000	1.000	1.000
School Re-open	0.948	1.009	1.025	1.000	1.248	0.995
Restaurant Closure	1.000	1.000	0.974	0.996	1.342	0.998
Restaurant Re-open	1.001	1.001	1.074	1.009	1.054	1.001

The proposed metric of the MAE ratio between factual and counterfactual has the limitation that if the intervention has a small effect on the progression of the disease, it would be nearly-optimal, close to 1, regardless of the quality of the counterfactual outcomes. To address, we also introduce an additional metric which measures how closely the forecasting model tracks the change in the simulated data due to the intervention, defined as the correlation between the prediction difference (counterfactual prediction - factual prediction) and the ground truth difference generated by the simulator (counterfactual ground truth - factual ground truth) across different locations. We evaluate this metric on various what-if scenarios and forecasting dates. Table 20 shows that in most scenarios and forecasting dates, the Pearson correlations between prediction differences are higher than 0.6 which represents strong correlations. In some what-if scenarios, the correlations are even higher than 0.9 such as School re-opening. Note that the correlations can be different across different forecasting dates because the factual scenarios can be different at different forecasting dates and the model can fit them differently.

Supplementary Table 20 | Pearson correlations between (counterfactual - factual prediction) and (counterfactual - factual ground truth) in 6 different scenarios.

Forecasting date	9/7		9/28		10/19	
Compartments	Confirmed	Deaths	Confirmed	Deaths	Confirmed	Deaths
Mobility Increase (10x)	0.667	0.578	0.693	0.716	0.692	0.627
Mobility Decrease (0.9x)	0.666	0.598	0.701	0.723	0.691	0.632
School Closure	0.459	0.813	0.888	0.797	0.664	0.558
School Re-open	0.526	0.610	0.943	0.531	0.948	0.924
Restaurant Closure	0.789	0.699	0.442	0.703	0.678	0.602
Restaurant Re-open	0.377	0.517	0.723	0.801	0.848	0.849

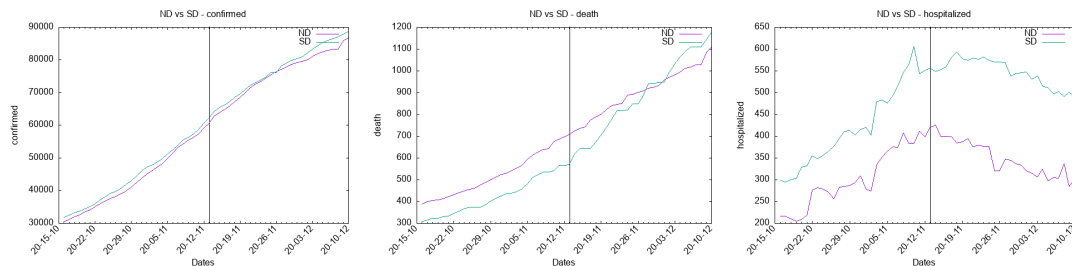
Matched pairs: Matching is a technique sometimes used in epidemiological studies to deduce causal effects from observational data[6, 7, 8, 9]. Matching method as an evaluation tool for our counterfactual analysis involves finding pairs of locations that are similar demographically and are in similar stages of the spread of COVID-19, but implement different NPI policies. The idea of matching is to eliminate the confounders - hidden variables other than the policies themselves that might be affecting the outcome.

On the US state-level data, we use an algorithm to find matched pairs of states based on similarity of confirmed cases, deaths, and mobility in the past 28 days, as well as similarity of their demographics (total population, population density, and income per capita). We looked for twins that implemented the same NPI policies in the past 28 days, but then their policies diverged. The most compelling matched pair of states we found were North Dakota and South Dakota. The two states have similar demographic metrics (see Table 21). They also had the same policies and COVID-19 metrics prior to 11/13/2020. After 11/13, North Dakota mandated mask wearing and closed restaurants, while South Dakota remained open (some counties in South Dakota introduced their own mask wearing mandates, but the state did not).

Supplementary Table 21 | Demographic metrics for North Dakota and South Dakota.

	North Dakota	South Dakota
Population	762062	884659
Population Density	4.265 ppl./sq. mi.	4.506 ppl./sq. mi.
Income Per Capita	\$36002.7	\$30873.1

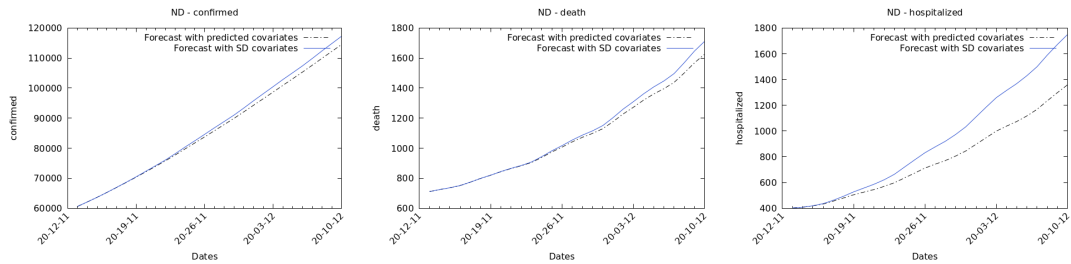
Figure 10 shows confirmed confirmed cases, deaths, and hospitalization in North Dakota and South Dakota in the 28 days prior and 28 days following the introduction of lockdown policies in North Dakota. While 28 days is insufficient to see the full impact of these policies, we can see confirmed cases in North Dakota increasing slower than in South Dakota after the stricter policies are mandated, and the same trend is visible in the plots of deaths and hospitalized cases.



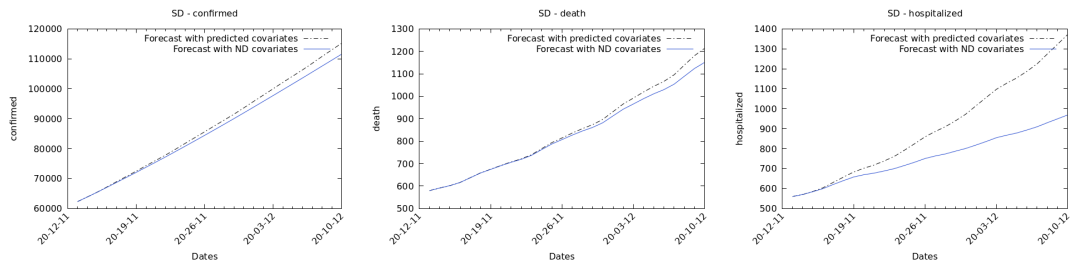
Supplementary Figure 10 | Confirmed cases, death, and hospitalizations in North Dakota (ND) vs. South Dakota (SD). The vertical line is the date on which ND introduced stricter lockdown policies.

Once a matched pair of states is found, it can be used to evaluate the validity of our counterfactual predictions. More specifically, we can pretend for the purposes of evaluation that South Dakota applied North Dakota's stricter policies, while North Dakota applied South Dakota's less strict policies.

Figure 11 shows the counterfactual outcomes when North Dakota overrides its forecasted NPIs with the actual NPIs of South Dakota. North Dakota's forecasted confirmed cases, deaths, and hospitalizations increase, mirroring the higher increase in predicted confirmed cases, deaths, and hospitalizations we observe in South Dakota. Figure 12 shows the reversed counterfactual outcomes – if we replace the South Dakota's forecasted NPIs with North Dakota's stricter NPIs, the forecasts of confirmed cases, deaths, and hospitalizations are lowered for South Dakota, mirroring more closely the trend in the ground truth values of North Dakota. Overall, this matched-pair study improves the confidence on the reliability of our counterfactual analysis framework.



Supplementary Figure 11 | Confirmed cases, death, and hospitalizations for ND with SD overrides



Supplementary Figure 12 | Confirmed cases, death, and hospitalizations for SD with ND overrides

Using past scenarios: A common way to apply counterfactuals in practice is to impart additional information about the future to the model. We may know that a large event scheduled for next week would dramatically increase mobility in a region, but the forecasting model has no way of predicting that this large event would occur. For example, the motorcycle rally on August 7th in Sturgis, SD in the USA caused a mobility increase not predicted by the baseline forecast generated on August 5th. This baseline forecast should predict cases and deaths less accurately than a counterfactual scenario on August 5th with (a more accurate) increased mobility. We could have told the model about the increased mobility as a counterfactual and ask it to adjust its forecast accordingly. We propose to evaluate the model by overriding the forecasted future features (such as NPIs and mobility) to make its predictions with the actual observed values of these features. This technique is particularly effective when unpredictable events happen and the observed values deviate far from the forecasts.

To identify past scenarios where the models' forecasted features deviated from the actual

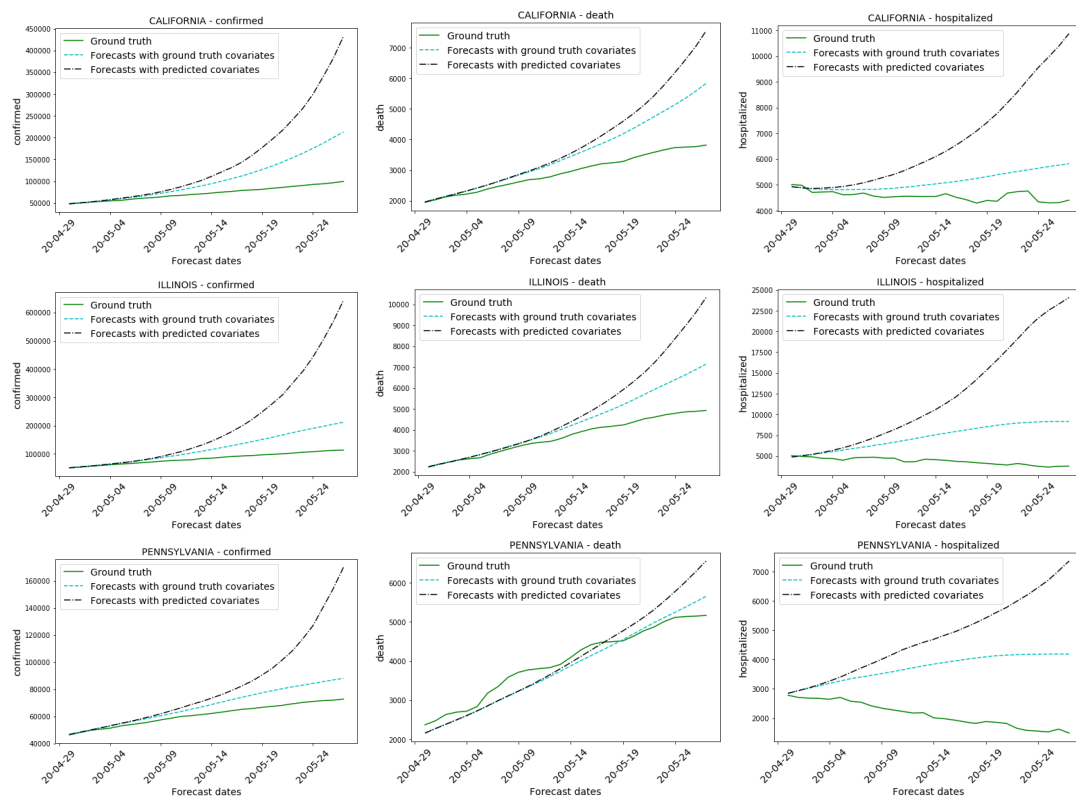
values, we train the feature forecasting component of our model on an expanding time window and analyze the average prediction errors. Figure 14 (left) illustrates the feature forecasting performance measured by MAE across different training window sizes; the feature forecasting performance improves with increasing training dataset size. We pick 6 forecasting dates for evaluation in the range from 2020-04 to 2020-10: 2 dates from 2020-05 with inaccurate feature forecasting, 1 date from 2020-07 with moderately accurate feature forecasting, and 3 more recent dates with accurate feature forecasting. We then evaluated how well our framework performed when forecasting deaths and confirmed cases in the future window of 28 days for each one of those scenarios.

We then correlated the improvements in the accuracy of our models' forecast of deaths and confirmed cases due to using the correct overrides with the accuracy of our feature forecasts. We measured the improvement in the deaths/confirmed cases forecast accuracy with the ratio between the MAPE of forecasts based on actual features and MAPE of forecasts based on predicted features (lower MAPE ratio corresponds to more improvement in accuracy). We show the Pearson correlation coefficient of the MAPE ratio and the MAE of feature forecasts in Figure 14 (right) and Table 22. We expect that the counterfactual overrides based on the true feature values would help the model more when the original feature forecasts it based its predictions on were far from the true values, as observed in Figure 14. Note the negative correlation between the MAPE ratio and the feature MAE, indicating that higher feature MAE (worse feature predictions), results in greater improvement in the accuracy of predicted deaths and confirmed cases (i.e. lower MAPE ratio) when the ground truth feature values are applied as the counterfactual overrides. The Pearson correlation coefficients between MAPE ratio and feature MAE are -0.998 and -0.934 for death and confirmed cases predictions, respectively.

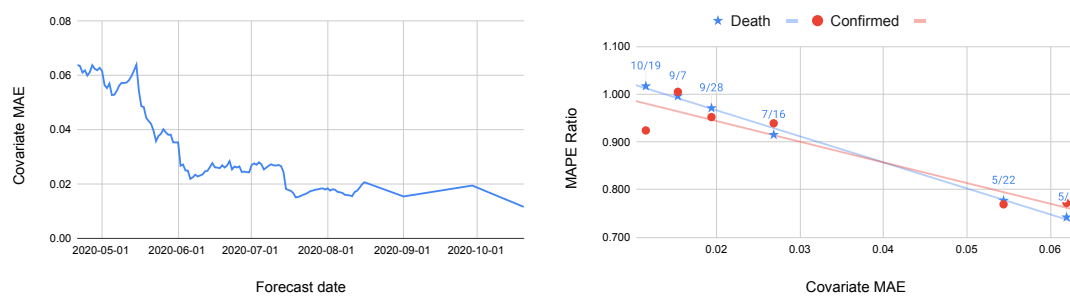
Figure 13 exemplifies forecasts with actual features and predicted features for three states (CA, IL, PA) for future 28 days in 2020-05. Forecasts move much closer to the ground truth when we override the forecasted features with the actual observed values.

Supplementary Table 22 | Forecasting performance improvements with historical ground truth feature. MAE/MAPE ratio = MAE/MAPE with ground truth feature / MAE with forecasted feature for the US model.

Forecasting date	MAE Ratio		MAPE ratio		MAE of feature forecasts
	Confirmed	Deaths	Confirmed	Deaths	
5/5	0.539	0.759	0.743	0.772	0.062
5/22	0.646	0.734	0.777	0.769	0.054
7/16	0.877	0.957	0.915	0.939	0.026
9/7	0.899	0.994	0.996	1.005	0.015
9/28	0.968	0.956	0.971	0.952	0.019
10/19	0.996	0.970	1.017	0.924	0.011



Supplementary Figure 13 | Confirmed cases, death, and hospitalization forecasts with actual features and predicted features in comparison to ground truth. Each row illustrates the forecasting results in California, Illinois, and Pennsylvania, respectively.



Supplementary Figure 14 | (left) feature forecasting MAE, (right) feature forecasting MAE vs. what-if metric in terms of the death and confirmed MAPE ratio of inference with ground truth features over inference with forecasted features.

Counterfactual Analysis on Applying Vaccinations Drives with NPIs

Vaccination drives are ongoing around the globe. In the US, states have begun vaccinations for certain classes of occupations deemed vulnerable or crucial to controlling the pandemic. Here we run counterfactual analysis on the effect of vaccination, in combination with removing or reapplying the interventions—the State of Emergency in Japan and the 7 individual NPIs in the US, on the progression of the disease.

For the US, we consider:

1. The three vaccination rates from Supplementary Table 24: 0.1% (low), 0.5% (medium), and 1% (high) of the population per day.
2. The Rand corporation study on NPIs and their health and economic impacts[10] defined 5 levels of NPIs. From those levels, we focus on Levels 1, 3 and 5 for our study. The Levels and their corresponding definitions are described in Supplementary Table 23.

For Japan, we consider:

1. Four vaccination scenarios from Supplementary Table 24—vaccination rates of 0.1% (low), 0.5% (medium), 1% (high) and 2% (very high) of the population per day—compared to the baseline (expected trend-rate in the US and 0% in Japan).
2. Two Non-Pharmaceutical Interventions (NPI) scenarios—the State of Emergency being in force for all prefectures, and being withdrawn in all prefectures.

The application of one or more NPIs has a range of impacts and is subject to socioeconomic and compliance costs. Using high-frequency data such as electricity consumption and nitrogen dioxide emissions, Demirgüç-Kunt et al[11] show that the application of NPIs imposed an economic cost. Countries that applied NPIs early saw less of an impact over the period of the study as compared to countries that delayed their NPIs. An analogous compliance cost for NPIs were studied by Kantor and Kantor[12] and Seale et al[13]. Cross-sectional surveys were conducted to understand self-reported compliance with NPIs and the perceived complexity or cost in such compliance. They showed that the perceived complexity had an impact on NPI compliance. For example, hand-washing was perceived to be less onerous than mask-wearing or school closures, hence there was more self-reported acceptance and compliance with hand-washing. Ghaffar zadegan[14] examines a different aspect of the application of NPIs. They show that applying multiple NPIs simultaneously is more beneficial than applying them individually. The nonlinear positive effects of multiple NPIs outweigh the costs incurred in their application and compliance.

We apply these scenarios both individually, and in combinations considering the combinatorial space of 15 scenarios for Japan, and 16 scenarios for the US (Supplementary Figure 15). One additional, 'heterogeneous' scenario[17] is considered where 4 selected prefectures or states receive very high rate vaccinations, and the remaining prefectures or states receive medium rate vaccinations. The selected prefectures in Japan are Tokyo, Kanagawa, Aichi and Osaka. The selected states in the US are California, Florida, New York and Texas.

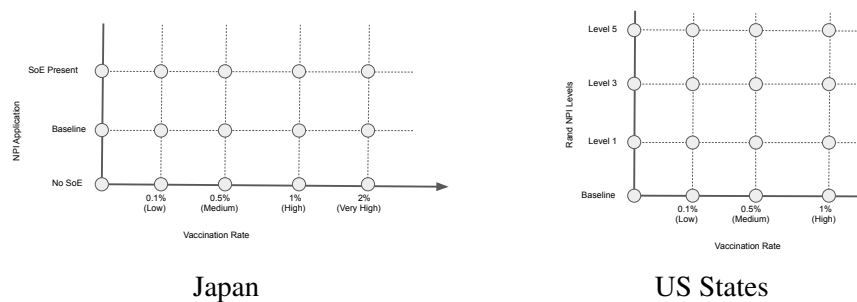
Here, our baseline scenario is the scenario where no counterfactuals are applied, and the forecasted features are used in the model to make predictions. For example, Supplementary

NPI	Level 1	Level 3	Level 5
Shut Bars/Restaurants	Remove	Impose	Impose
Shut Non-Essential Businesses	Remove	Impose	Impose
Restrict Movement (Shelter-in-Place)	Remove	Remove	Impose
Close Schools	Impose	Impose	Impose
Emergency Declaration	Impose	Impose	Impose
Restrict Gatherings	Remove	Impose	Impose
Require Masks	Baseline	Baseline	Baseline

Supplementary Table 23 | Selected subset of the NPI levels in US states as defined by[10]. The 'Baseline' setting means that the corresponding NPI is unchanged from its ground truth values

Counterfactual Vaccination Drive	Description
Low vaccination rate	0.1% of the population is vaccinated daily (95% effectiveness[15, 16])
Medium vaccination rate	0.5% of the population is vaccinated daily (95% effectiveness[15, 16])
High vaccination rate	1% of the population is vaccinated daily (95% effectiveness[15, 16])
Very high vaccination rate	2% of the population is vaccinated daily (95% effectiveness[15, 16])

Supplementary Table 24 | Counterfactual Vaccination Drive levels in Japan and US states. The 'Baseline' setting means that the corresponding NPI is unchanged from its ground truth values

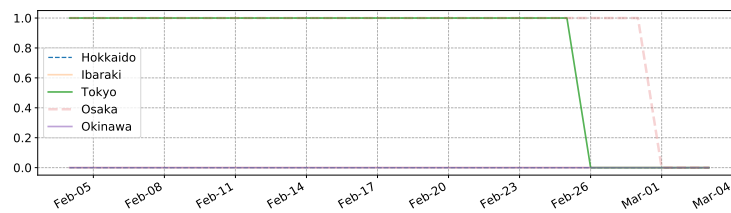


Supplementary Figure 15 | Considered potential scenarios for Japan and the US.

Figure 16 shows the forecasted State of Emergency feature for selected prefectures. In a prefecture, say Tokyo, where the State of Emergency is in force as of the training horizon of Feb. 4 2021, the forecasted feature could change value in the forecasting horizon.

Counterfactual Analysis Results for Japan

Supplementary Table 25 shows that vaccination drives by themselves may not be able to drive the disease to the disease-free equilibrium (all compartments tend towards 0 except for susceptible cases) after 28 days. The reduction in predicted exposed count ranges up to 28.4% in Ibaraki. However, in combination with NPIs, vaccination drives have bigger predicted impact, ranging from 75% to 76.5% in Ibaraki. The heterogeneous scenario, where very high vaccination rates are applied to Tokyo, Kanagawa, Aichi and Osaka, with the remaining prefectures receiving medium rates, show an overall national benefit of 2.1% fewer predicted exposed count.



Supplementary Figure 16 | Forecasted State of Emergency feature for selected prefectures. Observe that for those prefectures where the current value is 1.0, the forecasted value does not remain at 1.0 in the forecasting horizon (Feb. 4 2021 through Mar. 3 2021).

Vaccination Drives	Baseline	State of Emergency OFF	State of Emergency ON
Baseline (same 28-day trend)	Osaka: 5779	Osaka: 9853 (70.5%)	Osaka: 5652 (-2.2%)
	Hyogo: 4894	Hyogo: 11840 (141.95%)	Hyogo: 4453 (-9.01%)
	Okinawa: 3838	Okinawa: 3838 (0.0%)	Okinawa: 2262 (-41.06%)
	Hokkaido: 3253	Hokkaido: 3253 (0.0%)	Hokkaido: 2203 (-32.28%)
	Ibaraki: 3210	Ibaraki: 3210 (0.0%)	Ibaraki: 791 (-75.36%)
Low (0.1%)	Osaka: 5772 (-0.12%)	Osaka: 9781 (69.27%)	Osaka: 5648 (-2.27%)
	Hyogo: 4828 (-1.35%)	Hyogo: 11618 (137.39%)	Hyogo: 4399 (-10.11%)
	Okinawa: 3814 (-0.63%)	Okinawa: 3814 (-0.63%)	Okinawa: 2261 (-41.09%)
	Hokkaido: 3237 (-0.49%)	Hokkaido: 3237 (-0.49%)	Hokkaido: 2202 (-32.31%)
	Ibaraki: 3161 (-1.53%)	Ibaraki: 3161 (-1.53%)	Ibaraki: 789 (-75.42%)
Medium (0.5% population/day)	Osaka: 5744 (-0.59%)	Osaka: 9498 (64.37%)	Osaka: 5632 (-2.53%)
	Hyogo: 4568 (-6.66%)	Hyogo: 10748 (119.62%)	Hyogo: 4185 (-14.47%)
	Okinawa: 3717 (-3.15%)	Okinawa: 3717 (-3.15%)	Okinawa: 2256 (-41.22%)
	Hokkaido: 3174 (-2.43%)	Hokkaido: 3174 (-2.43%)	Hokkaido: 2197 (-32.43%)
	Ibaraki: 2969 (-7.51%)	Ibaraki: 2969 (-7.51%)	Ibaraki: 782 (-75.64%)
High (1.0%)	Osaka: 5710 (-1.19%)	Osaka: 9150 (58.33%)	Osaka: 5613 (-2.87%)
	Hyogo: 4249 (-13.16%)	Hyogo: 9705 (98.3%)	Hyogo: 3923 (-19.84%)
	Okinawa: 3598 (-6.25%)	Okinawa: 3598 (-6.25%)	Okinawa: 2249 (-41.4%)
	Hokkaido: 3096 (-4.83%)	Hokkaido: 3096 (-4.83%)	Hokkaido: 2192 (-32.62%)
	Ibaraki: 2737 (-14.74%)	Ibaraki: 2737 (-14.74%)	Ibaraki: 773 (-75.92%)
Very high (2.0%)	Osaka: 5641 (-2.37%)	Osaka: 8474 (46.63%)	Osaka: 5574 (-3.55%)
	Hyogo: 3634 (-25.75%)	Hyogo: 7763 (58.62%)	Hyogo: 3414 (-30.24%)
	Okinawa: 3365 (-12.32%)	Okinawa: 3365 (-12.32%)	Okinawa: 2236 (-41.74%)
	Hokkaido: 2941 (-9.59%)	Hokkaido: 2941 (-9.59%)	Hokkaido: 2181 (-32.92%)
	Ibaraki: 2299 (-28.38%)	Ibaraki: 2299 (-28.38%)	Ibaraki: 755 (-76.48%)
Heterogeneous scenario	Overall Benefit: -11436 vs -11203 (Medium rate vaccination and State of Emergency)		
	Osaka: -3.55%		
	Hyogo: -14.47%		
	Okinawa: -41.22%		
	Hokkaido: -32.44%		
	Ibaraki: -75.65%		

Supplementary Table 25 | Counterfactual number of predicted exposed count on the 28th-day of the forecasting horizon for the top 5 prefectures by confirmed cases. Lower numbers are better outcomes. Negative percentages imply a reduction in counts, positive percentages imply an increase.

Counterfactual Analysis Results for the US

For the US model, the NPIs imposed by state governments are more granular and act on different aspects of public life, unlike the Japan model, which only has the binary-valued State of Emergency series. As described previously, we use 3 different levels of NPIs. We present counterfactual analysis on the effect of each scenario on predicted exposed count in Supplementary Table 26. As an example, in California, the reduction benefit in the number of predicted exposed count can be as much as 33% reductions when high-rate vaccination drives are applied in tandem with imposing Level 5 NPIs universally. Of this total benefit, 23% is due to the vaccination drive (the reduction in exposed count in California between the 'Consistent' scenario and the 'High rate/Consistent NPI' scenario). A further 10% reduction in predicted exposed counts occurs from the 'High rate/Consistent NPI' scenario to the 'High rate/Level 5 NPI' scenario.

Vaccination Drives	Baseline	Rand Level 1	Rand Level 3	Rand Level 5
Baseline (same 28-day trend)	Florida: 7085317 Texas: 3644234 California: 3320749 New York: 2383868 Georgia: 2365059	Florida: 5503106 (-22.33%) Texas: 3379252 (-7.27%) California: 3939997 (18.65%) New York: 2430069 (1.94%) Georgia: 1823232 (-22.91%)	Florida: 4295254 (-39.38%) Texas: 3090946 (-15.18%) California: 2555856 (-23.03%) New York: 2290595 (-3.91%) Georgia: 1361138 (-42.45%)	Florida: 3953503 (-44.2%) Texas: 2923793 (-19.77%) California: 2226157 (-32.96%) New York: 2070954 (-13.13%) Georgia: 1271457 (-46.24%)
Low (0.1%)	Florida: 6955601 (-1.83%) Texas: 3442140 (-5.55%) California: 3234428 (-2.6%) New York: 2198528 (-7.77%) Georgia: 2329769 (-1.49%)	Florida: 5448184 (-23.11%) Texas: 3260573 (-10.53%) California: 3804434 (14.57%) New York: 2225056 (-6.66%) Georgia: 1807612 (-23.57%)	Florida: 4283798 (-39.54%) Texas: 3042372 (-16.52%) California: 2532294 (-23.74%) New York: 2157782 (-9.48%) Georgia: 1358821 (-42.55%)	Florida: 3951766 (-44.23%) Texas: 2911744 (-20.1%) California: 2222985 (-33.06%) New York: 2040171 (-14.42%) Georgia: 1271135 (-46.25%)
Medium (0.5%)	Florida: 6444873 (-9.04%) Texas: 3022900 (-17.05%) California: 2897022 (-12.76%) New York: 2037002 (-14.55%) Georgia: 2190393 (-7.39%)	Florida: 5229632 (-26.19%) Texas: 2983522 (-18.13%) California: 3277258 (-1.31%) New York: 2050938 (-13.97%) Georgia: 1745359 (-26.2%)	Florida: 4238021 (-40.19%) Texas: 2918479 (-19.92%) California: 2438532 (-26.57%) New York: 2022620 (-15.15%) Georgia: 1349559 (-42.94%)	Florida: 3944822 (-44.32%) Texas: 2879254 (-20.99%) California: 2210308 (-33.44%) New York: 1996826 (-16.24%) Georgia: 1269847 (-46.31%)
High (1.0%)	Florida: 5824653 (-17.79%) Texas: 2935710 (-19.44%) California: 2556876 (-23.0%) New York: 2004098 (-15.93%) Georgia: 2020171 (-14.58%)	Florida: 4958994 (-30.01%) Texas: 2924629 (-19.75%) California: 2780332 (-16.27%) New York: 2010821 (-15.65%) Georgia: 1668052 (-29.47%)	Florida: 4180890 (-40.99%) Texas: 2892860 (-20.62%) California: 2335454 (-29.67%) New York: 1997826 (-16.19%) Georgia: 1337992 (-43.43%)	Florida: 3936144 (-44.45%) Texas: 2870274 (-21.24%) California: 2195779 (-33.88%) New York: 1985910 (-16.69%) Georgia: 1268238 (-46.38%)
Heterogeneous scenario	Overall Benefit: 29685292 vs 29776868 (Medium rate vaccination and Level 5 NPIs) Florida: 3919800 (-44.68%) Texas: 2859707 (-21.53%) California: 2183781 (-34.24%) New York: 1976346 (-17.09%) Georgia: 1269847 (-46.31%)			

Supplementary Table 26 | Counterfactual analysis for the predicted exposed count on the 28th-day of the forecasting horizon for the top 5 states. Negative percentages imply a reduction in counts, positive percentages imply an increase.

Key observations

From examining results from counterfactual simulations in US states and Japanese prefectures, we can draw certain initial observations:

1. Performing vaccinations, even at high rates, does not appreciably reduce the predicted number of predicted confirmed cases over the forecasting horizon of 28 days. The change in predicted deaths is even smaller after 28 days. However, reductions are observed in the predicted exposed and susceptible counts.

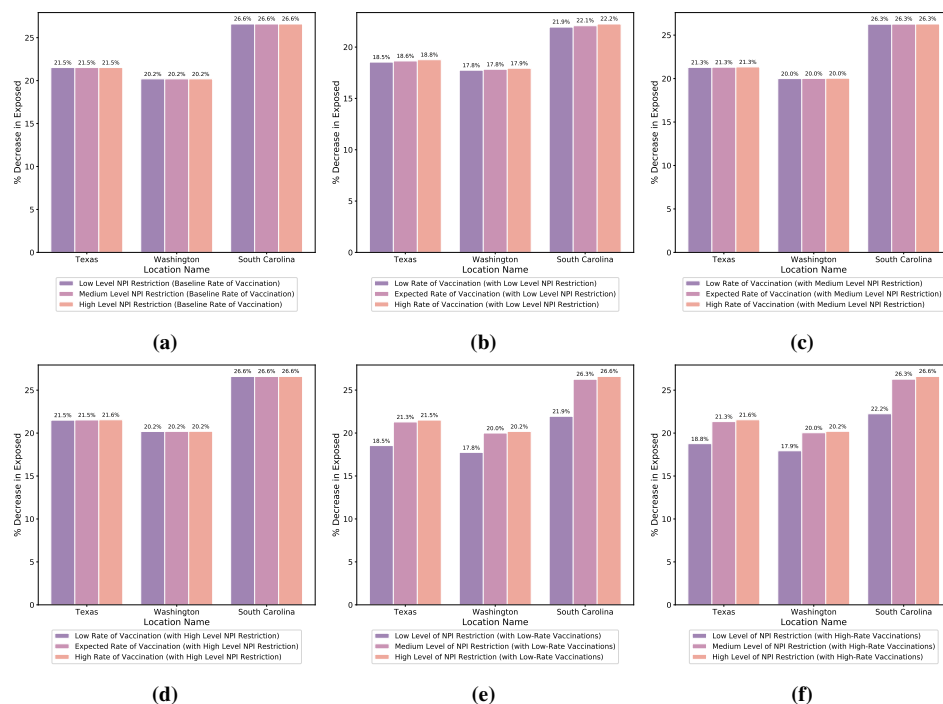
State Name	Scenario	Counts
Texas	Forecasted Features Baseline	2582026
Virginia	Forecasted Features Baseline	667035
California	Forecasted Features Baseline	598596
Ohio	Forecasted Features Baseline	563736
New York	Forecasted Features Baseline	469758
Texas	Decrease Mobility by 90%	2571370 (-0.41)
Virginia	Decrease Mobility by 90%	665642 (-0.21)
California	Decrease Mobility by 90%	597829 (-0.13)
Ohio	Decrease Mobility by 90%	562366 (-0.24)
New York	Decrease Mobility by 90%	469361 (-0.08)
Texas	Decrease Mobility by 90% after 7 days	2574306 (-0.3)
Virginia	Decrease Mobility by 90% after 7 days	666060 (-0.15)
California	Decrease Mobility by 90% after 7 days	598094 (-0.08)
Ohio	Decrease Mobility by 90% after 7 days	562744 (-0.18)
New York	Decrease Mobility by 90% after 7 days	469526 (-0.05)
Texas	All NPIs in force	1844165 (-28.58)
Virginia	All NPIs in force	544914 (-18.31)
California	All NPIs in force	525464 (-12.22)
Ohio	All NPIs in force	447292 (-20.66)
New York	All NPIs in force	441189 (-6.08)
Texas	All NPIs in force after 7 days	2018082 (-21.84)
Virginia	All NPIs in force after 7 days	581784 (-12.78)
California	All NPIs in force after 7 days	549385 (-8.22)
Ohio	All NPIs in force after 7 days	477844 (-15.24)
New York	All NPIs in force after 7 days	452624 (-3.65)

Supplementary Table 27 | Effect of delaying the application of mobility restrictions and NPIs in the US. Shown are the counterfactual outcomes for the predicted exposed count after the 28-day forecasting horizon for the top 5 US states ordered by baseline confirmed cases. Percent change from the forecasted features baseline are shown in parentheses. Negative percentages imply a reduction in counts, whereas positive percentages imply an increase.

2. Keeping NPIs seems to play a crucial role during vaccination drives. The reduction in predicted confirmed cases is significant when the NPIs are kept along with vaccination efforts. The reductions in the predicted exposed or susceptible counts is larger.
3. The heterogeneous application of very high rate vaccinations applied to the more populated states, with medium rate vaccinations in the remaining states can result in fewer predicted cases and deaths.

location_name	Scenario	Counts
California	Forecasted Features Baseline	14173271
Texas	Forecasted Features Baseline	8749567
Pennsylvania	Forecasted Features Baseline	7342288
Michigan	Forecasted Features Baseline	6656776
Georgia	Forecasted Features Baseline	6170734
California	Decrease Mobility by 90%	14174349 (0.01)
Texas	Decrease Mobility by 90%	8763605 (0.16)
Pennsylvania	Decrease Mobility by 90%	7343064 (0.01)
Michigan	Decrease Mobility by 90%	6657352 (0.01)
Georgia	Decrease Mobility by 90%	6171784 (0.02)
California	Decrease Mobility by 90% after 7 days	14173908 (0.0)
Texas	Decrease Mobility by 90% after 7 days	8758933 (0.11)
Pennsylvania	Decrease Mobility by 90% after 7 days	7342741 (0.01)
Michigan	Decrease Mobility by 90% after 7 days	6657123 (0.01)
Georgia	Decrease Mobility by 90% after 7 days	6171300 (0.01)
California	All NPIs in force	14278178 (0.74)
Texas	All NPIs in force	9742591 (11.35)
Pennsylvania	All NPIs in force	7397400 (0.75)
Michigan	All NPIs in force	6692778 (0.54)
Georgia	All NPIs in force	6229372 (0.95)
California	All NPIs in force after 7 days	14236135 (0.44)
Texas	All NPIs in force after 7 days	9447979 (7.98)
Pennsylvania	All NPIs in force after 7 days	7373938 (0.43)
Michigan	All NPIs in force after 7 days	6679216 (0.34)
Georgia	All NPIs in force after 7 days	6203820 (0.54)

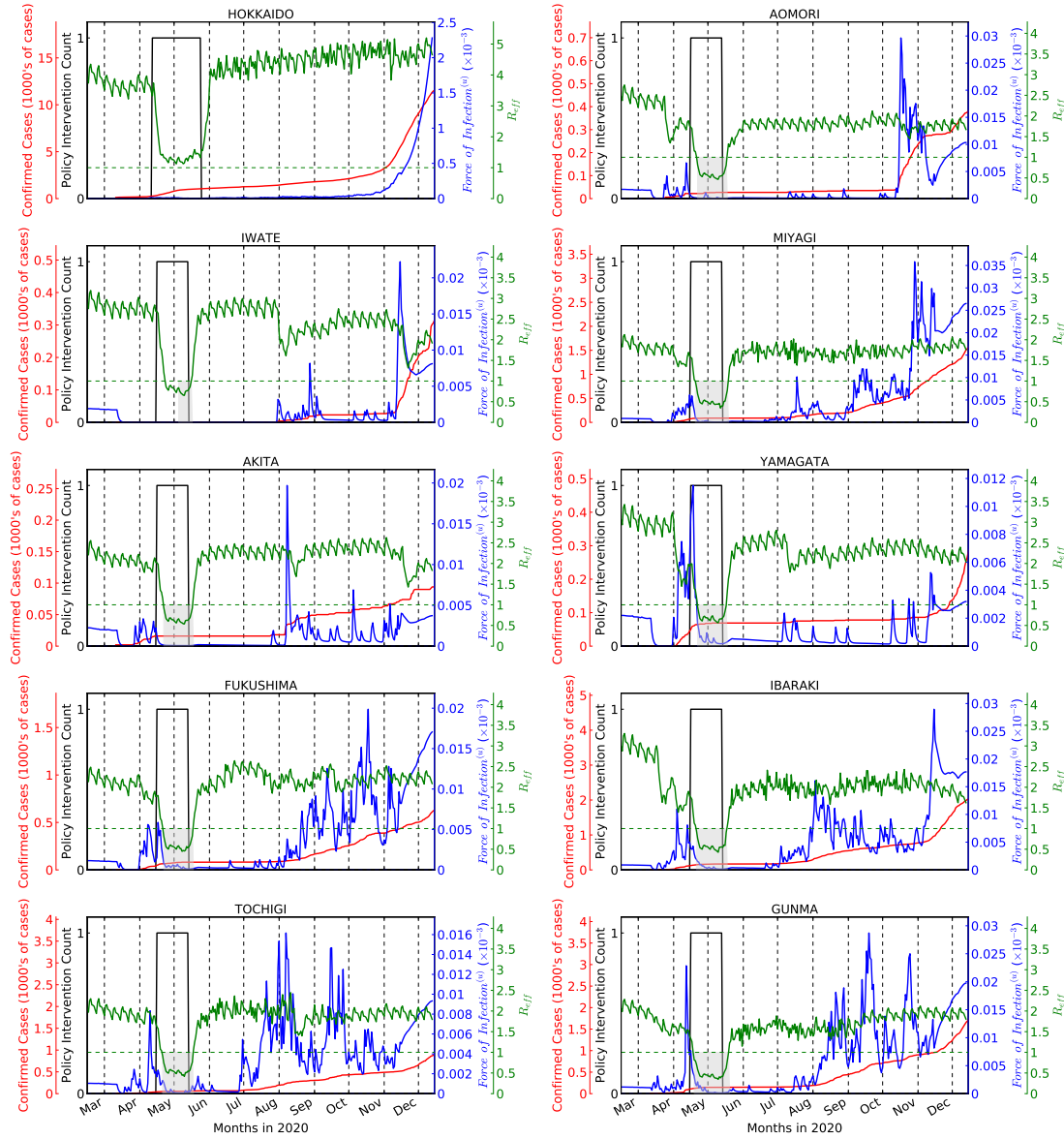
Supplementary Table 28 | Effect of delaying the application of mobility restrictions and NPIs in the US. Shown are the counterfactual number of predicted susceptible counts after the 28-day forecasting horizon for the top 5 US states ordered by baseline confirmed cases. Percent change from the forecasted features baseline are shown in parentheses. Negative percentages imply a reduction in counts, whereas positive percentages imply an increase.



Supplementary Figure 17 | Counterfactual analysis on the predicted exposed counts with different vaccination rates for 3 selected US states (Supplementary Table 24) in tandem with NPIs (Supplementary Table 23), for the prediction date of March 1, 2021. Here, the baseline exposed counts are 69694, 67591 and 63742 for Texas, Washington and South Carolina, respectively. The percent decrease in predicted exposed counts compared to the baseline, are shown for the counterfactual cases that (a) the NPI Levels are changed over 1-Low, 3-Medium and 5-High, while the vaccination rates are the expected forecasted rates; (b) the NPI Levels are at 1-Low, while the vaccination rates are varied; (c) the NPI Levels are at 3-Medium, while the vaccination rates are varied; (d) the NPI Levels are at 5-High, while the vaccination rates are varied; (e) the NPI Levels are changed over 1-Low, 3-Medium and 5-High, while the vaccination rates are low; (f) the NPI Levels are changed over 1-Low, 3-Medium and 5-High, while the vaccination rates are high.

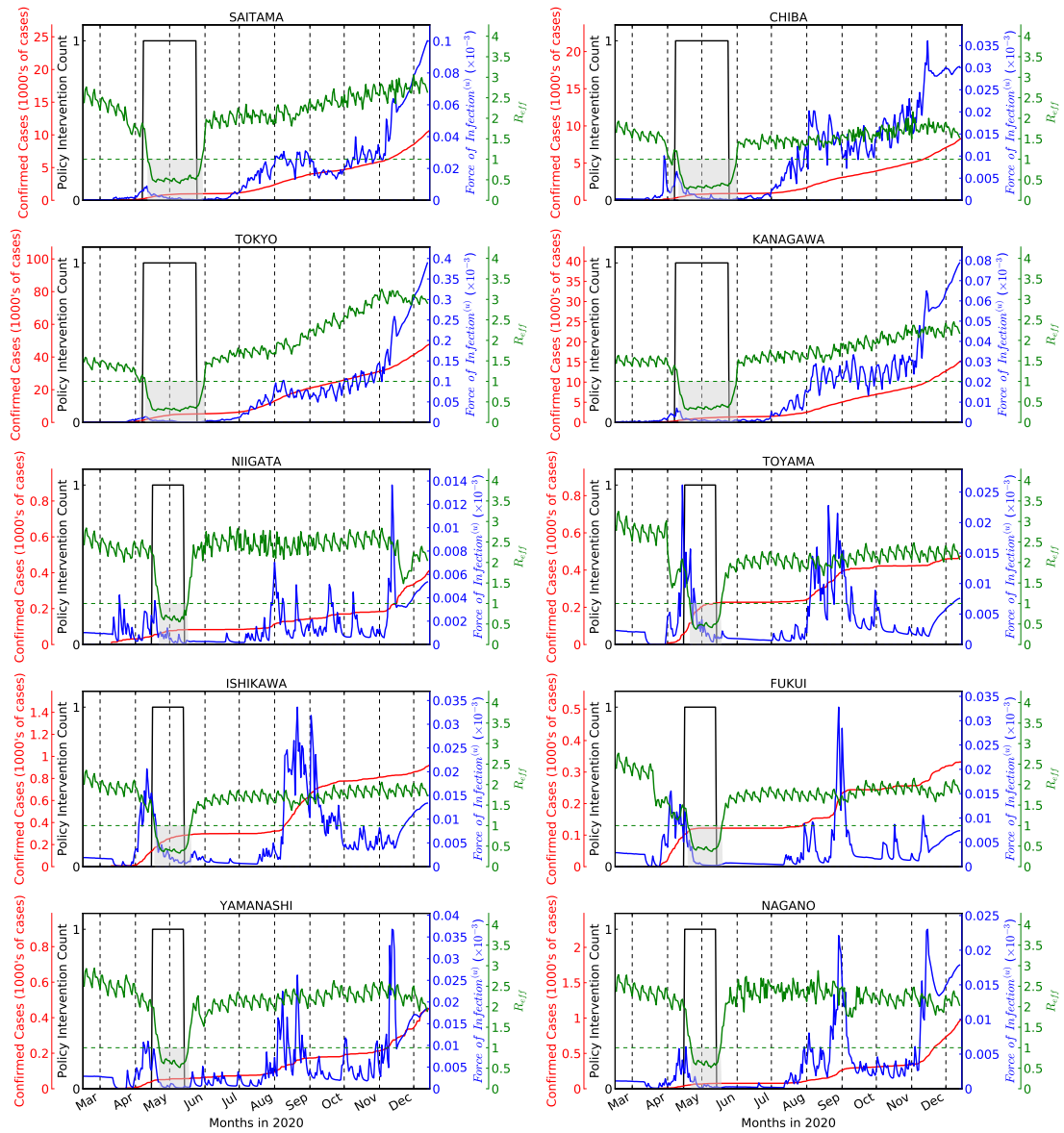
4 Additional explainability results

Figures 18 through 22 show the progression of the disease in Japanese prefectures, along with the decrease in the Force of Infection F_u and R_{eff} when NPIs are applied. The Force of Infection spikes on certain dates in a handful of prefectures: Akita (JP-05), Fukui (JP-18), Tottori (JP-31) and Shimane (JP-32). Those dates also correspond to a sharp spike in the predicted confirmed cases. This could indicate data restatements in those prefectures.



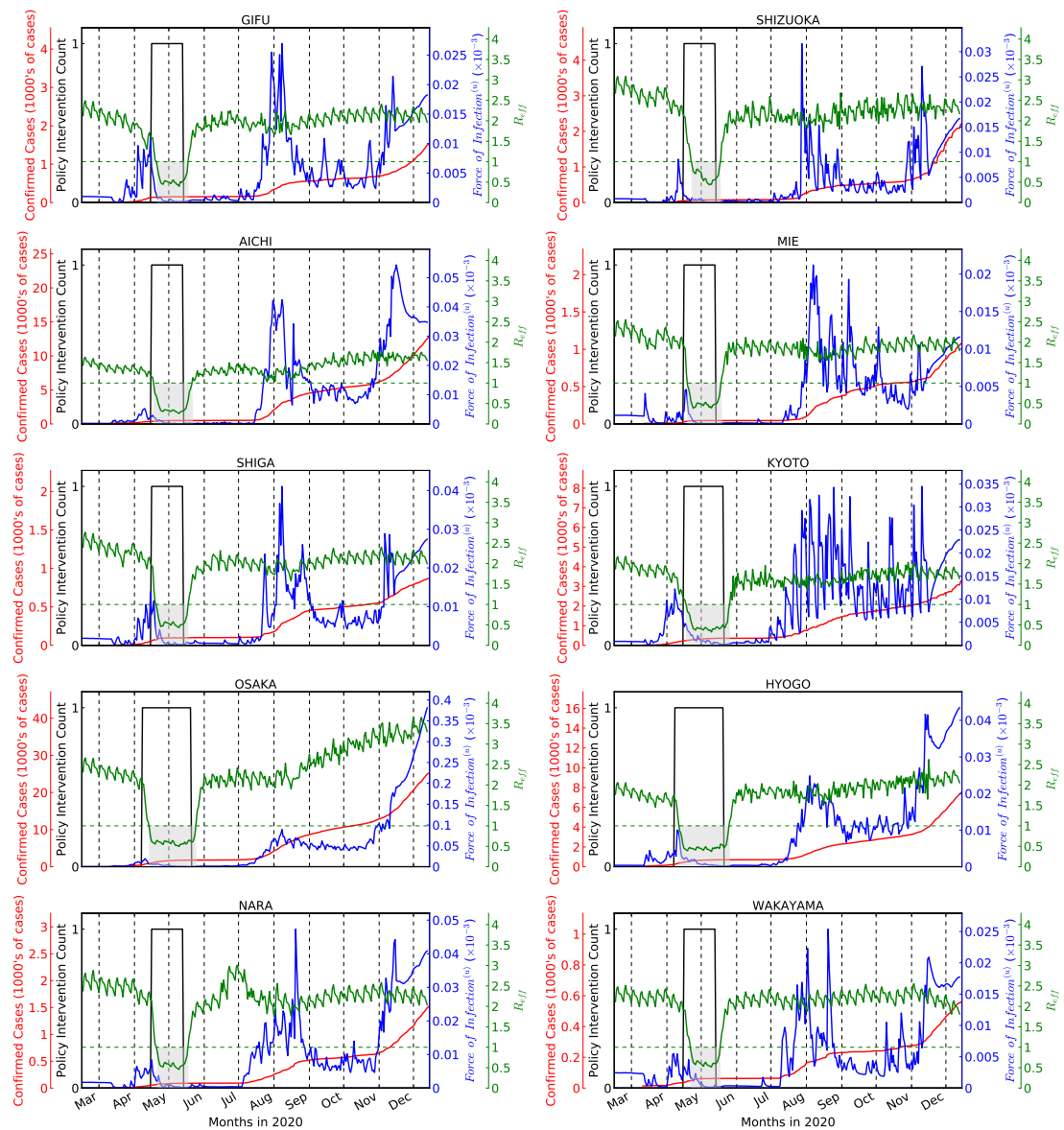
Supplementary Figure 18 | Progression of the disease in Japanese Prefectures (JP-01 through JP-10)

Figures 23 through 28 show the progression of the disease in US states and territories, along

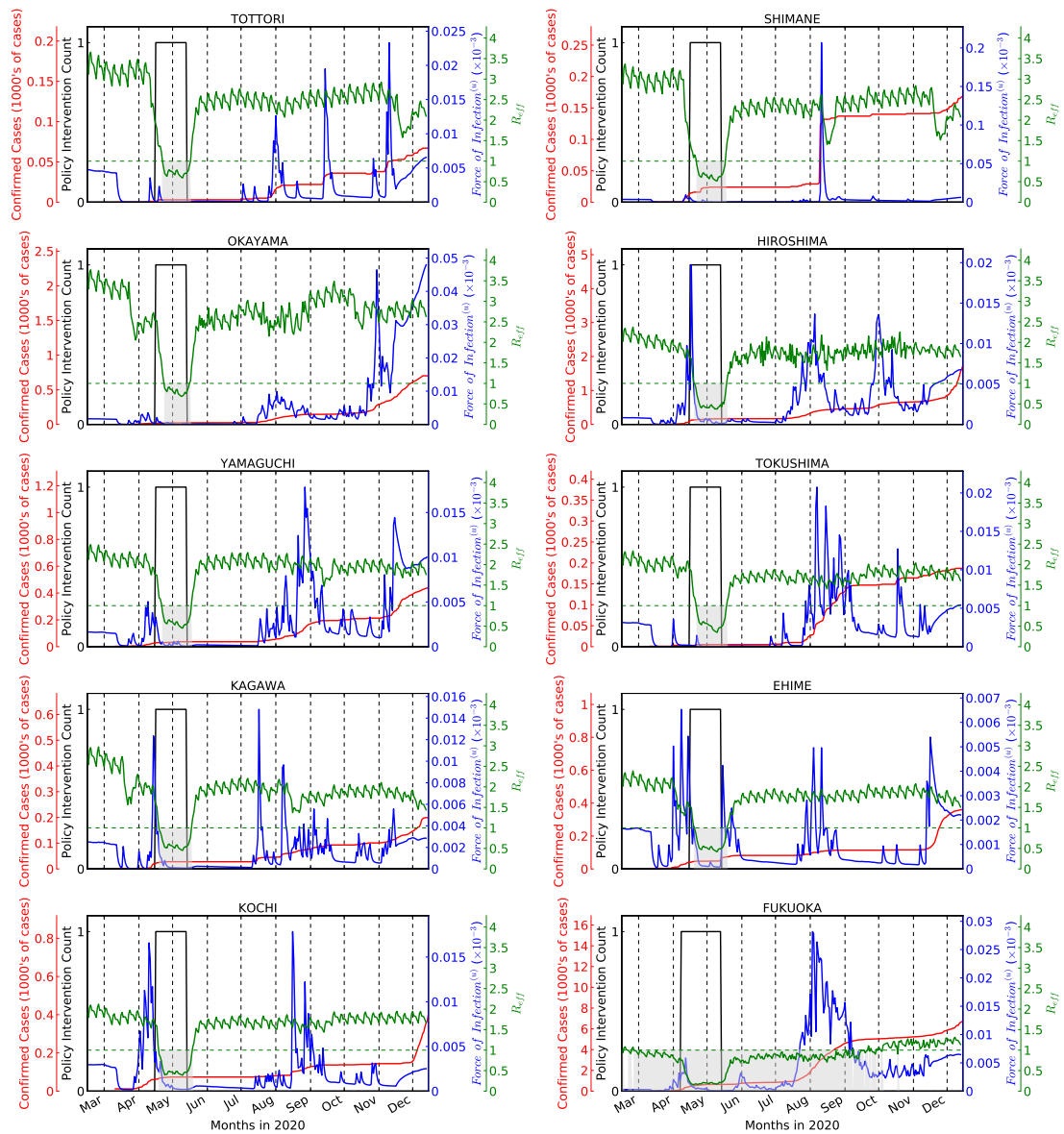


Supplementary Figure 19 | Progression of the disease in Japanese Prefectures (JP-11 through JP-20)

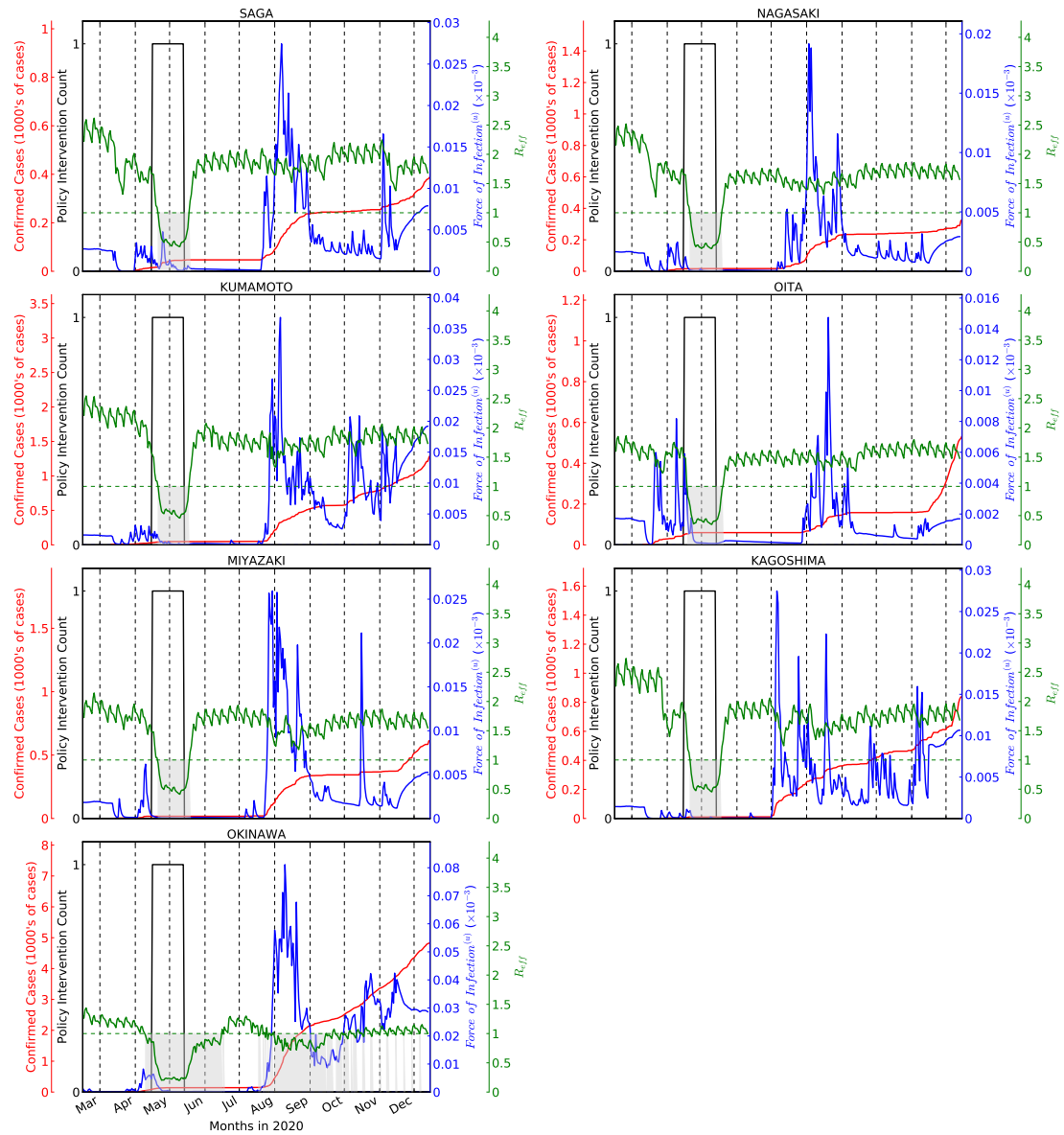
with the decrease in the Force of Infection F_u and R_{eff} when NPIs are applied.



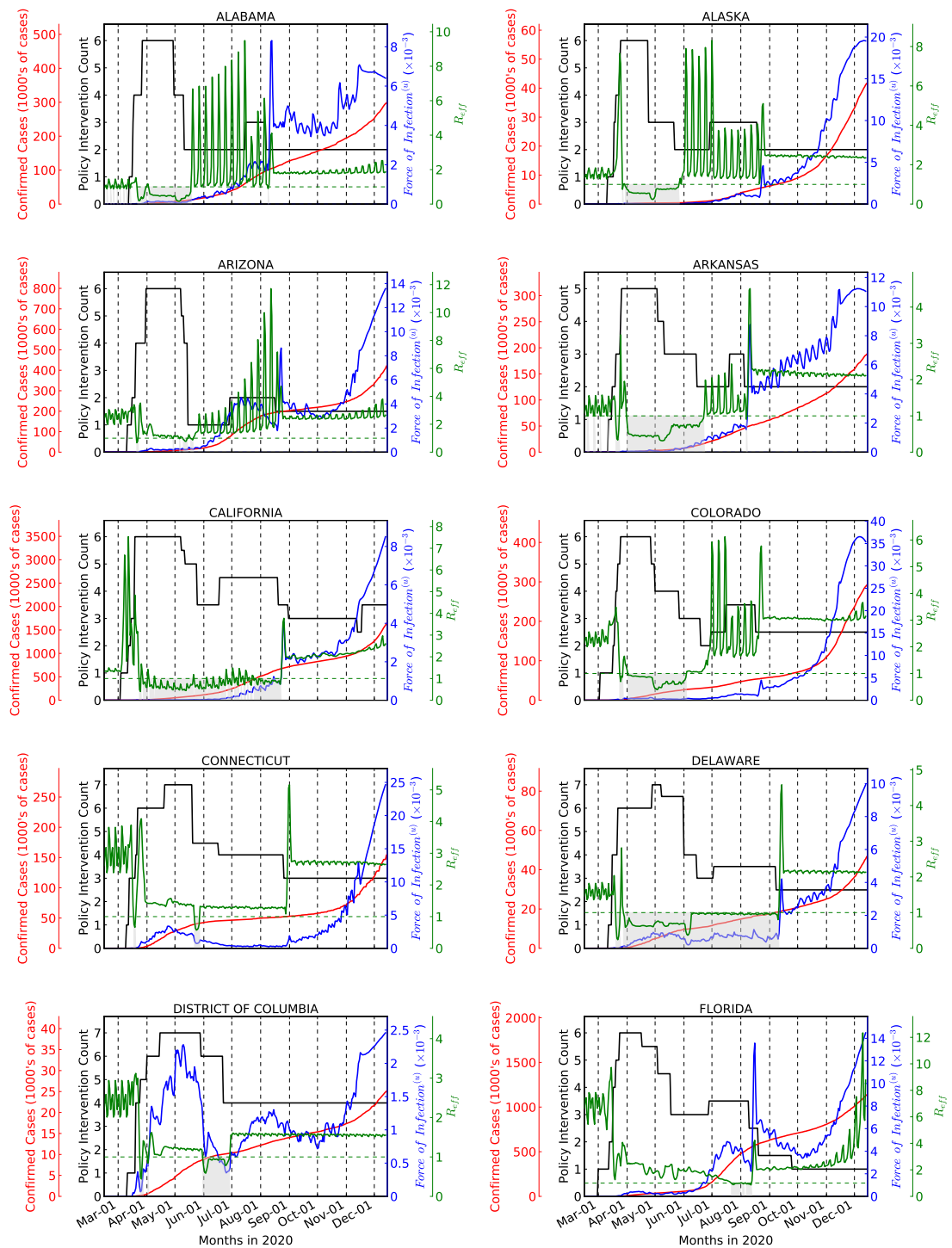
Supplementary Figure 20 | Progression of the disease in Japanese Prefectures (JP-21 through JP-30)



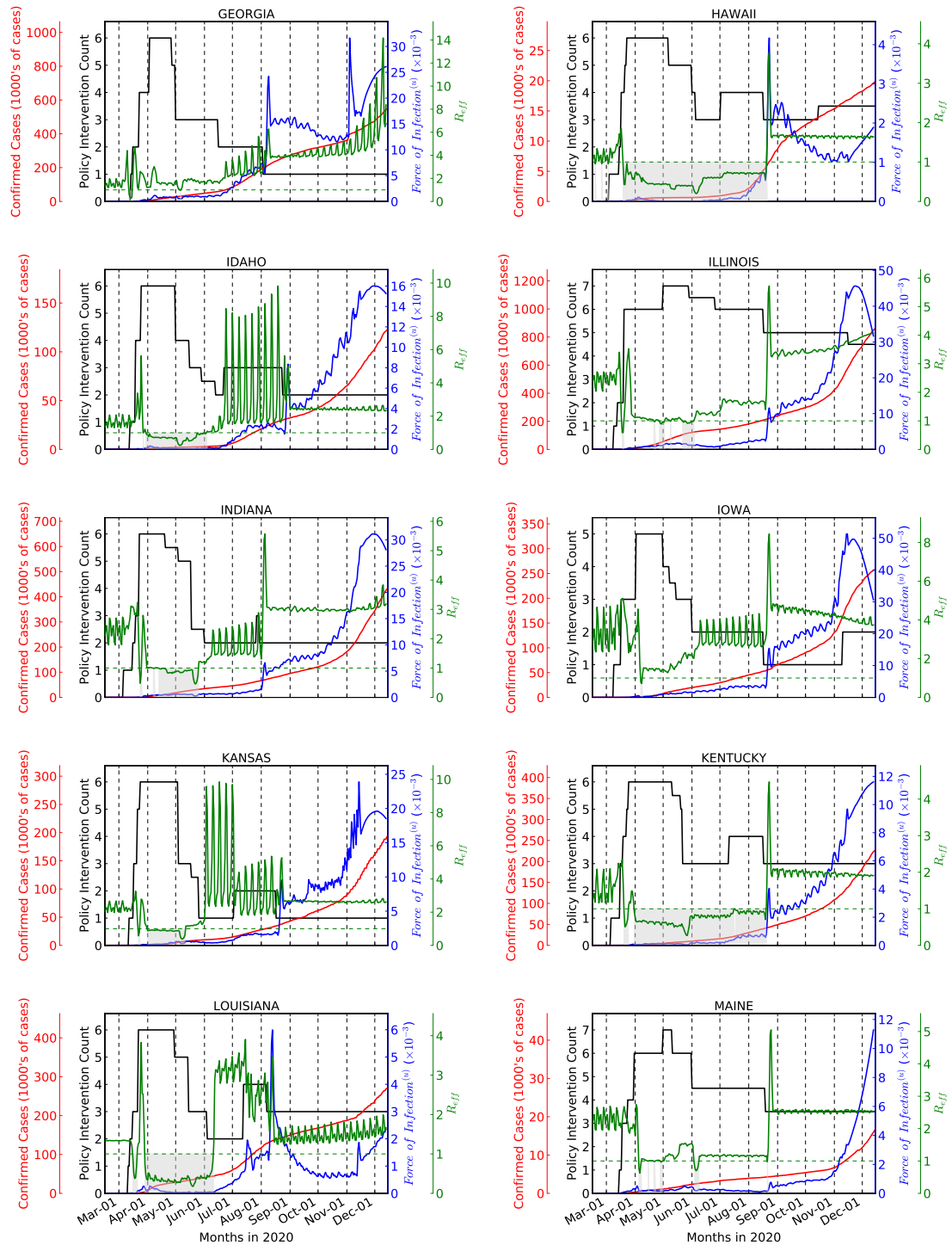
Supplementary Figure 21 | Progression of the disease in Japanese Prefectures (JP-31 through JP-40)



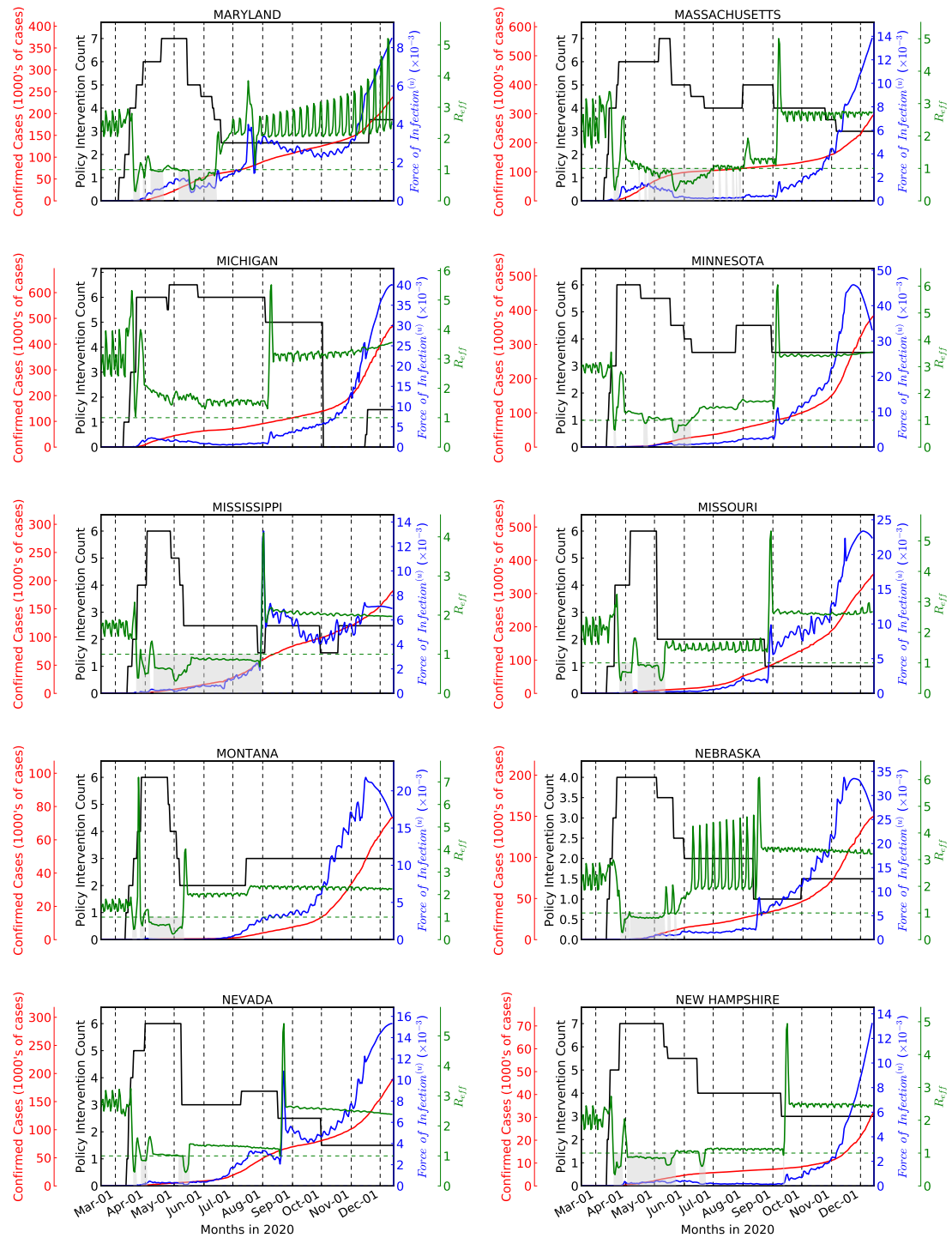
Supplementary Figure 22 | Progression of the disease in Japanese Prefectures (JP-41 through JP-47)



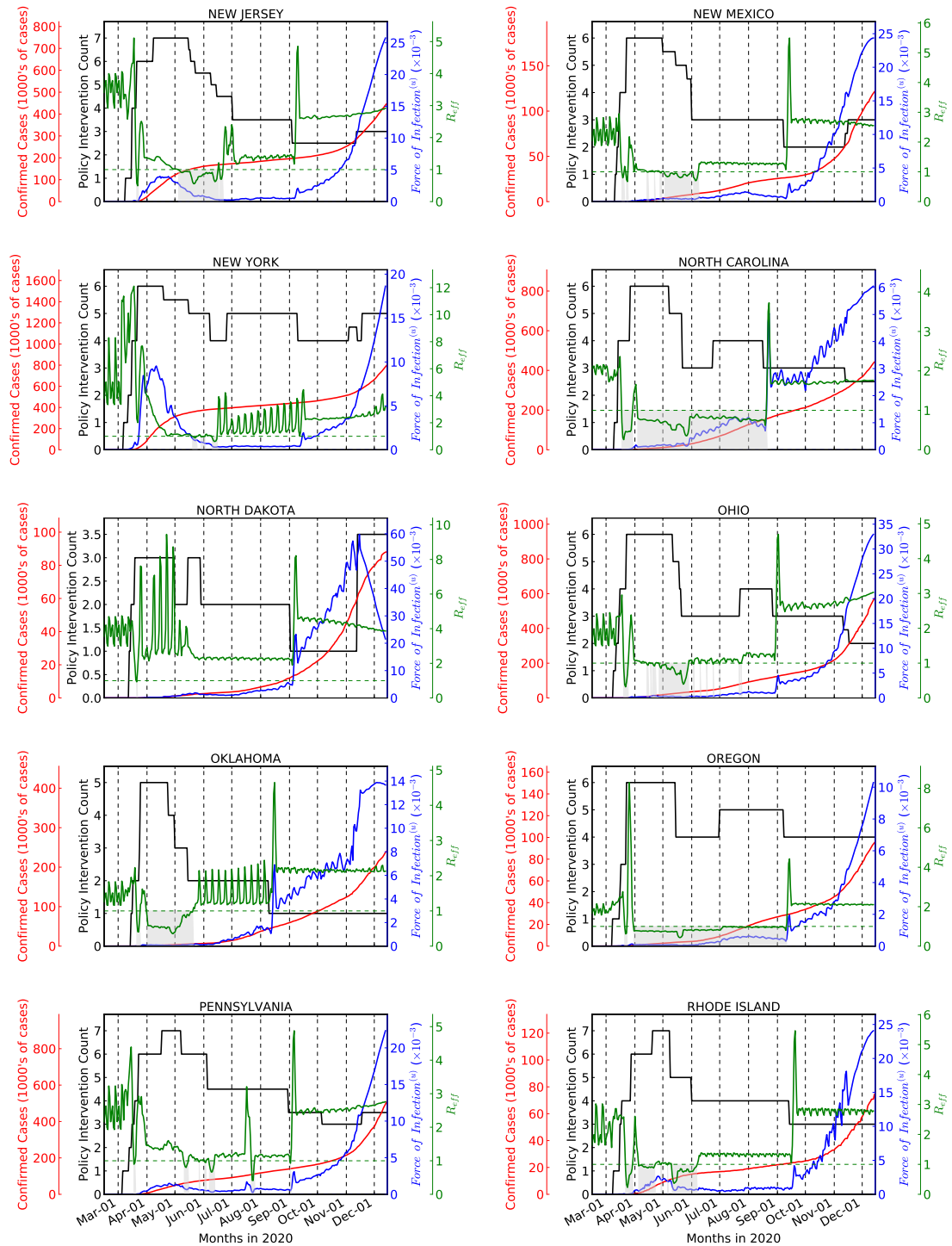
Supplementary Figure 23 | Progression of the disease in US States (Alabama through Florida)



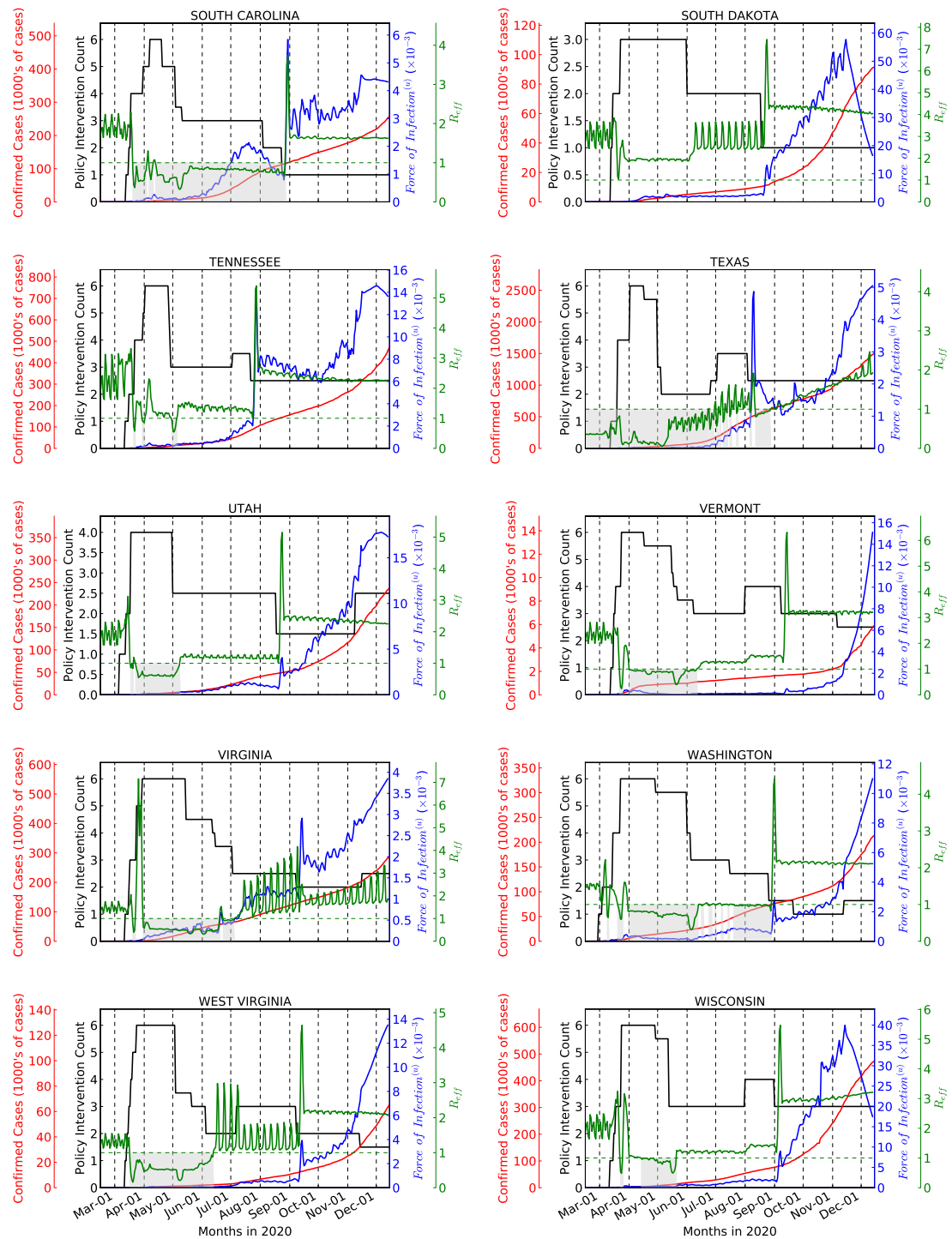
Supplementary Figure 24 | Progression of the disease in US States (Georgia through Maine)



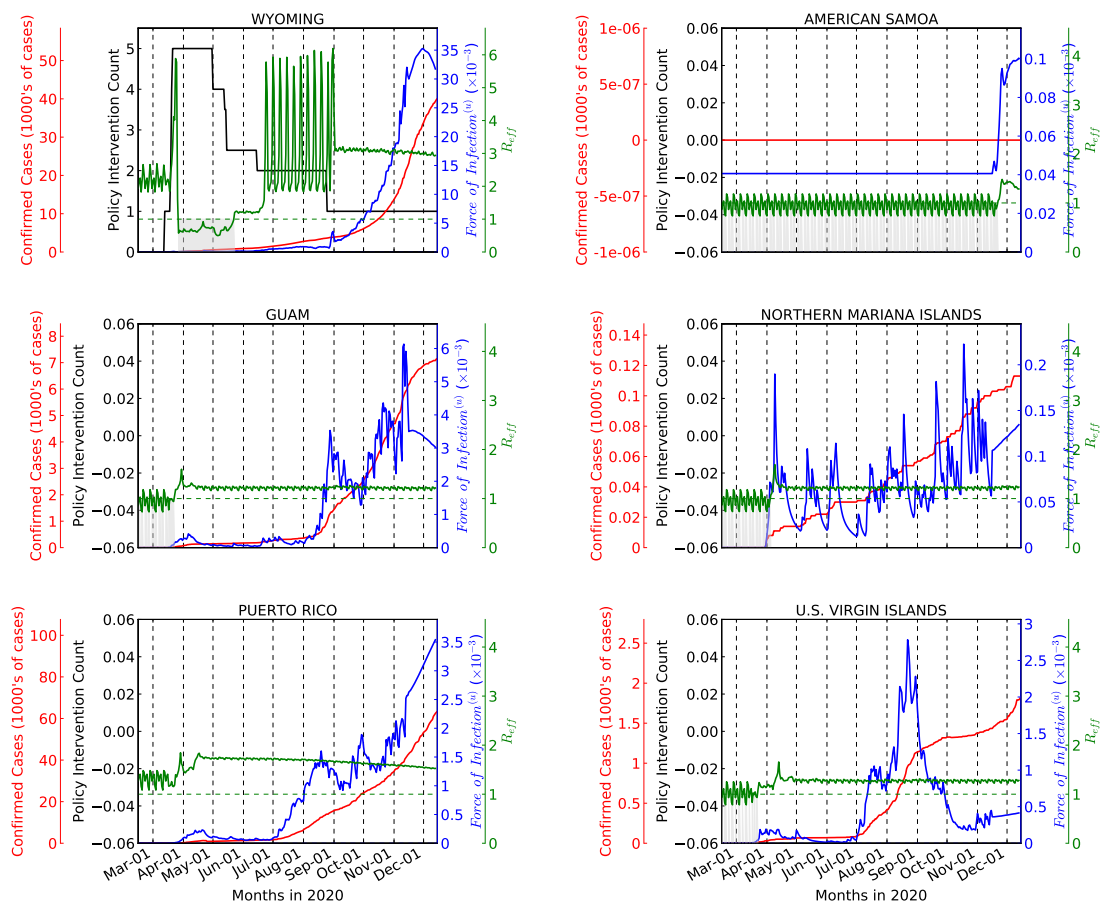
Supplementary Figure 25 | Progression of the disease in US States (Maryland through New Hamsphire)



Supplementary Figure 26 | Progression of the disease in US States (New Jersey through Rhode Island)



Supplementary Figure 27 | Progression of the disease in US States (South Carolina through Wisconsin)



Supplementary Figure 28 | Progression of the disease in US States (Wyoming through U.S. Virgin Islands)

Feature Name	Median Rank
Number of Households	1
Per Capita Income	2
Mean Air Quality Index	3
Ratio of Population over 60	4
Population Density	5

For Average Documented Contact Rate

Feature Name	Median Rank
Average Rating	1
Per Capita Income	2
Mean Air Quality Index	3
Number of Critical Access Hospitals	4
Ratio of Population over 60	5
Households on Public Assistance	6
Population Density	7
Average Patient Experience	8
Emergency Services Supported	9
Number of ICU Beds	10
Number of Acute Care Hospitals	11
Non-Emergency Services Supported	12
Number of Households	13
Above Average Patient Experience	14
Below Average Patient Experience	15

For Hospitalization Rate

Feature Name	Median Rank
Ratio of Population over 60	1
Number of Critical Access Hospitals	2
Per Capita Income	3
Below Average Patient Experience	4
Above Average Patient Experience	5
Non-Emergency Services Supported	6
Average Patient Experience	7
Mean Air Quality Index	8
Emergency Services Supported	9
Households on Public Assistance	10
Population Density	11
Average Rating	12
Number of Acute Care Hospitals	13
Number of Households	14
Number of ICU Beds	15

For ICU Admission Rate

Feature Name	Median Rank
Above Average Patient Experience	1
Emergency Services Supported	2
Per Capita Income	3
Number of Acute Care Hospitals	4
Ratio of Population over 60	5
Number of Critical Access Hospitals	6
Number of Households	7
Below Average Patient Experience	8
Average Patient Experience	9
Non-Emergency Services Supported	10
Average Rating	11
Households on Public Assistance	12
Population Density	13

For Documented Recovery Rate

Feature Name	Median Rank
Non-Emergency Services Supported	1
Average Patient Experience	2
Above Average Patient Experience	3
Number of Acute Care Hospitals	4
Emergency Services Supported	5
Ratio of Population over 60	6
Number of Households	7
Population Density	8
Below Average Patient Experience	9
Number of Critical Access Hospitals	10
Average Rating	11
Households on Public Assistance	12
Per Capita Income	13

For Recovery Rate from Hospitals

Feature Name	Median Rank
Ratio of Population over 60	1
Per Capita Income	2
Mean Air Quality Index	3
Population Density	4
Number of Households	5

For Average Undocumented Contact Rate

Feature Name	Median Rank
Number of Critical Access Hospitals	1
Average Rating	2
Ratio of Population over 60	3
Non-Emergency Services Supported	4
Below Average Patient Experience	5
Households on Public Assistance	6
Number of Households	7
Number of Acute Care Hospitals	8
Average Patient Experience	9
Above Average Patient Experience	10
Population Density	11
Mean Air Quality Index	12
Emergency Services Supported	13
Per Capita Income	14

For Reinfection Rate

Feature Name	Median Rank
Population Density	1
Number of Critical Access Hospitals	2
Households on Public Assistance	3
Per Capita Income	4
Emergency Services Supported	5
Average Patient Experience	6
Number of Acute Care Hospitals	7
Above Average Patient Experience	8
Average Rating	9
Number of ICU Beds	10
Number of Households	11
Non-Emergency Services Supported	12
Below Average Patient Experience	13
Ratio of Population over 60	14
Mean Air Quality Index	15

For Ventilator Rate

Feature Name	Median Rank
Below Average Patient Experience	1
Non-Emergency Services Supported	2
Number of Households	3
Number of Acute Care Hospitals	4
Average Patient Experience	5
Above Average Patient Experience	6
Emergency Services Supported	7
Population Density	8
Number of Critical Access Hospitals	9
Average Rating	10
Per Capita Income	11
Ratio of Population over 60	12
Households on Public Assistance	13

For Undocumented Recovery Rate

Feature Name	Median Rank
Population Density	1
Households on Public Assistance	2
Average Patient Experience	3
Average Rating	4
Non-Emergency Services Supported	5
Number of Acute Care Hospitals	6
Per Capita Income	7
Ratio of Population over 60	8
Emergency Services Supported	9
Number of Critical Access Hospitals	10
Number of Households	11
Below Average Patient Experience	12
Above Average Patient Experience	13

For Recovery Rate from ICUs

Supplementary Figure 29 | Ranking of static features for encoders in the US Model of January 20, 2021 (Part 1).

Feature Name	Median Rank
Population Density	1
Number of Critical Access Hospitals	2
Households on Public Assistance	3
Per Capita Income	4
Average Rating	5
Average Patient Experience	6
Ratio of Population over 60	7
Non-Emergency Services Supported	8
Emergency Services Supported	9
Number of Acute Care Hospitals	10
Number of Households	11
Above Average Patient Experience	12
Below Average Patient Experience	13

For Recovery Rate from Ventilators

Feature Name	Median Rank
Population Density	1
Per Capita Income	2
Average Rating	3
Ratio of Population over 60	4
Households on Public Assistance	5
Mean Air Quality Index	6
Average Patient Experience	7
Number of Acute Care Hospitals	8
Non-Emergency Services Supported	9
Number of Households	10
Below Average Patient Experience	11
Above Average Patient Experience	12
Emergency Services Supported	13
Number of Critical Access Hospitals	14
Number of ICU Beds	15

For Documented Death Rate

Feature Name	Median Rank
Non-Emergency Services Supported	1
Number of ICU Beds	2
Number of Acute Care Hospitals	3
Average Patient Experience	4
Emergency Services Supported	5
Number of Households	6
Below Average Patient Experience	7
Number of Critical Access Hospitals	8
Ratio of Population over 60	9
Above Average Patient Experience	10
Mean Air Quality Index	11
Households on Public Assistance	12
Population Density	13
Per Capita Income	14
Average Rating	15

For Death Rate from Hospitals

Feature Name	Median Rank
Number of Critical Access Hospitals	1
Below Average Patient Experience	2
Number of ICU Beds	3
Mean Air Quality Index	4
Emergency Services Supported	5
Number of Households	6
Non-Emergency Services Supported	7
Number of Acute Care Hospitals	8
Population Density	9
Average Patient Experience	10
Ratio of Population over 60	11
Households on Public Assistance	12
Above Average Patient Experience	13
Per Capita Income	14
Average Rating	15

For Death Rate from ICUs

Feature Name	Median Rank
Population Density	1
Number of Critical Access Hospitals	2
Average Rating	3
Per Capita Income	4
Above Average Patient Experience	5
Below Average Patient Experience	6
Households on Public Assistance	7
Mean Air Quality Index	8
Average Patient Experience	9
Ratio of Population over 60	10
Emergency Services Supported	11
Number of ICU Beds	12
Number of Households	13
Number of Acute Care Hospitals	14
Non-Emergency Services Supported	15

For Death Rate from Ventilators

Feature Name	Median Rank
Ratio of Population over 60	1
Population Density	2
Per Capita Income	3
Number of Households	4
Average Rating	5

For Diagnosis Rate

Supplementary Figure 30 | Ranking of static features for encoders in the US Model of January 20, 2021 (Part 2).

Feature Name	Median Rank
NPI Schools	1
Deaths	2
Cases/Total Tests	3
Mobility Index	4
Confirmed Cases	5
Snowfall (mm)	6
Negative Antibody Test Ratio	7
NPI Non-Essential Business	8
NPI Bar/Restaurants	9
Positive Antibody Test Ratio	10
Cases Mean to Sum Ratio	11
Rainfall (mm)	12
Deaths Mean to Sum Ratio	13
Mobility Samples	14
NPI Gatherings	15
NPI Mask	16
Average Temperature (C)	17
NPI Movement	18
NPI Emergency	19

For Average Documented Contact Rate

Feature Name	Median Rank
NPI Schools	1
NPI Bar/Restaurants	2
Snowfall (mm)	3
Mobility Index	4
Cases/Total Tests	5
NPI Non-Essential Business	6
Mobility Samples	7
Average Temperature (C)	8
Deaths	9
Cases Mean to Sum Ratio	10
Deaths Mean to Sum Ratio	11
Confirmed Cases	12
NPI Movement	13
Positive Antibody Test Ratio	14
Negative Antibody Test Ratio	15
NPI Gatherings	16
NPI Mask	17
NPI Emergency	18
Rainfall (mm)	19

For Average Undocumented Contact Rate

Feature Name	Median Rank
Symptoms Search for Cough	1
Cases/Total Tests	2
Symptoms Search for Anosmia	3
Symptoms Search for Fever	4
Average Temperature (C)	5
Total Per Capita Tests	6
Snowfall (mm)	7
Positive Antibody Test Ratio	8
Day of Week	10
Symptoms Search for Chest Pain	10
Deaths Mean to Sum Ratio	11
Symptoms Search for Infection	11
Rainfall (mm)	12
Deaths	13
Confirmed Cases	14
Symptoms Search for Chills	16
Negative Antibody Test Ratio	17
Symptoms Search for Shortness of Breath	18
Cases Mean to Sum Ratio	19

For Hospitalization Rate

Feature Name	Median Rank
Snowfall (mm)	1
Symptoms Search for Anosmia	2
Symptoms Search for Chest Pain	3
Rainfall (mm)	4
Symptoms Search for Fever	6
Symptoms Search for Chills	7
Symptoms Search for Infection	8
Symptoms Search for Shortness of Breath	8
Total Per Capita Tests	9
Cases/Total Tests	10
Average Temperature (C)	11
Cases Mean to Sum Ratio	11
Symptoms Search for Cough	11
Deaths	13
Day of Week	15
Positive Antibody Test Ratio	15
Confirmed Cases	17
Deaths Mean to Sum Ratio	18
Negative Antibody Test Ratio	19

For Diagnosis Rate

Feature Name	Median Rank
Deaths	1
Confirmed Cases	2
Deaths Mean to Sum Ratio	3
Day of Week	4
Cases Mean to Sum Ratio	5

For ICU Admission Rate

Feature Name	Median Rank
Deaths Mean to Sum Ratio	1
Confirmed Cases	2
Cases Mean to Sum Ratio	3
Day of Week	4
Deaths	5

For Ventilator Rate

Feature Name	Median Rank
Rainfall (mm)	1
Deaths	2
Average Temperature (C)	3
Snowfall (mm)	4
Confirmed Cases	5
Cases Mean to Sum Ratio	6
Positive Antibody Test Ratio	7
Deaths Mean to Sum Ratio	8
Day of Week	9
Negative Antibody Test Ratio	10

For Documented Recovery Rate

Feature Name	Median Rank
Rainfall (mm)	1
Snowfall (mm)	2
Confirmed Cases	3
Average Temperature (C)	4
Deaths Mean to Sum Ratio	5
Positive Antibody Test Ratio	6
Deaths	7
Cases Mean to Sum Ratio	8
Negative Antibody Test Ratio	9

For Undocumented Recovery Rate

Feature Name	Median Rank
Day of Week	1
Deaths	2
Deaths Mean to Sum Ratio	3
Cases Mean to Sum Ratio	4
Confirmed Cases	5

For Recovery Rate from Hospitals

Feature Name	Median Rank
Confirmed Cases	1
Deaths Mean to Sum Ratio	2
Cases Mean to Sum Ratio	3
Deaths	4
Day of Week	5

For Recovery Rate from ICUs

Supplementary Figure 31 | Ranking of time-varying features for encoders for Encoders in the US Model of January 20, 2021 (Part 1).

Feature Name	Median Rank	Feature Name	Median Rank
Confirmed Cases	1	Cases Mean to Sum Ratio	1
Deaths	2	Deaths Mean to Sum Ratio	2
Deaths Mean to Sum Ratio	3	Day of Week	3
Cases Mean to Sum Ratio	4	Deaths	4
Day of Week	5	Confirmed Cases	5

For Recovery Rate from Ventilators Feature Ranks for Documented Death Rate

Feature Name	Median Rank	Feature Name	Median Rank
Day of Week	1	Day of Week	1
Confirmed Cases	2	Deaths	2
Deaths	3	Confirmed Cases	3
Deaths Mean to Sum Ratio	4	Deaths Mean to Sum Ratio	4
Cases Mean to Sum Ratio	5	Cases Mean to Sum Ratio	5

For Death Rate from Hospitals For Death Rate from ICUs

Feature Name	Median Rank
Day of Week	1
Confirmed Cases	2
Deaths Mean to Sum Ratio	3
Deaths	4
Cases Mean to Sum Ratio	5

For Death Rate from Ventilators

Supplementary Figure 32 | Ranking of time-varying features for encoders for Encoders in the US Model of January 20, 2021 (Part 2).

Feature Name	Median Rank
% Population that drink alcohol	1
BMI of Females (lower confidence interval)	2
People with H1N1 in 2010	3
Population over 75	4
Population Density	5
Average BMI of males	6
Per Capita Income	7
Population over 64	8
Population aged 0-14	9
Population aged 15-64	10
% Female Smokers	11
% Male Smokers	11

For Average Documented Contact Rate

Feature Name	Median Rank
Population over 75	1
% Population that drink alcohol	2
Population Density	3
Population aged 0-14	4
Population over 64	5
Population aged 15-64	6
Per Capita Income	7
Average BMI of males	8
BMI of Females (lower confidence interval)	9
People with H1N1 in 2010	10
% Female Smokers	11
% Male Smokers	11

For Reinfection Rate

Feature Name	Median Rank
Number of hospital beds	1
Population over 64	2
Population aged 15-64	3
Number of doctors	4
Population aged 0-14	5
Number of new ICU beds	6
Per Capita Income	7
Average BMI of males	8
Number of doctors/100k population	9
% Population that drink alcohol	10
Population over 75	11
Number of clinic beds	12
BMI of Females (lower confidence interval)	13
Number of clinic beds/100k population	14
People with H1N1 in 2010	15
Population Density	16
Number of hospital beds/100k population	17
% Female Smokers	18
% Male Smokers	18

For Documented Recovery Rate

Feature Name	Median Rank
Average BMI of males	1
People with H1N1 in 2010	2
Number of hospital beds/100k population	3
BMI of Females (lower confidence interval)	4
Number of doctors/100k population	5
Number of new ICU beds	6
Per Capita Income	7
Population aged 15-64	8
Population over 64	9
Number of clinic beds/100k population	10
Number of clinic beds	11
Population aged 0-14	12
Number of hospital beds	13
Number of doctors	14
Population over 75	15
% Population that drink alcohol	16
Population Density	17
% Female Smokers	18
% Male Smokers	18

For Recovery Rate from Hospitals

Feature Name	Median Rank
Number of clinic beds/100k population	1
Population Density	2
Population over 75	3
Number of new ICU beds	4
Number of doctors	5
Population aged 0-14	6
Number of doctors/100k population	7
Per Capita Income	8
% Population that drink alcohol	9
Population aged 15-64	10
Population over 64	11
Number of hospital beds/100k population	12
Number of hospital beds	13
Average BMI of males	14
BMI of Females (lower confidence interval)	15
People with H1N1 in 2010	16
Number of clinic beds	17
% Female Smokers	18
% Male Smokers	18

For Death Rate from Hospitals

Feature Name	Median Rank
People with H1N1 in 2010	1
BMI of Females (lower confidence interval)	2
% Population that drink alcohol	3
Average BMI of males	4
Population Density	5
Population over 64	6
Population aged 15-64	7
Per Capita Income	8
Population aged 0-14	9
Population over 75	10
% Female Smokers	11
% Male Smokers	11

For Average Undocumented Contact Rate

Feature Name	Median Rank
Number of clinic beds	1
Number of hospital beds	2
Number of new ICU beds	3
Number of doctors	4
Number of doctors/100k population	5
Number of clinic beds/100k population	6
Number of hospital beds/100k population	7

For Diagnosis Rate

Feature Name	Median Rank
Average BMI of males	1
People with H1N1 in 2010	2
Number of new ICU beds	3
Number of clinic beds/100k population	4
Number of doctors/100k population	5
BMI of Females (lower confidence interval)	6
Number of hospital beds/100k population	7
Population aged 0-14	8
Per Capita Income	9
Population over 64	10
Population aged 15-64	11
Number of doctors	12
Population Density	13
% Population that drink alcohol	14
Population over 75	15
Number of hospital beds	16
Number of clinic beds	17
% Female Smokers	18
% Male Smokers	18

For Undocumented Recovery Rate

Feature Name	Median Rank
Population over 75	1
Number of clinic beds/100k population	2
Number of new ICU beds	3
% Population that drink alcohol	4
People with H1N1 in 2010	5
Number of doctors/100k population	6
BMI of Females (lower confidence interval)	7
Number of clinic beds	8
Population Density	9
Per Capita Income	10
Average BMI of males	11
Number of hospital beds	12
Population aged 15-64	13
Number of hospital beds/100k population	14
Number of doctors	15
Population aged 0-14	16
Population over 64	17
% Female Smokers	18
% Male Smokers	18

For Documented Death Rate

Feature Name	Median Rank
Population over 75	1
Number of new ICU beds	2
People with H1N1 in 2010	3
Population aged 0-14	4
BMI of Females (lower confidence interval)	5
Number of clinic beds/100k population	6
Population aged 15-64	7
Number of doctors	8
Number of hospital beds	9
Per Capita Income	10
Population over 64	11
% Population that drink alcohol	12
Population Density	13
Number of clinic beds	14
Average BMI of males	15
Number of hospital beds/100k population	16
Number of doctors/100k population	17
% Female Smokers	18
% Male Smokers	18

For Hospitalization Rate

Feature Name	Median Rank
Mobility Changes: Residences	1
R_{eff}	2
Cases Mean to Sum Ratio	3
Deaths Mean to Sum Ratio	4
Weighted % of respondents reporting CLI	5
State of Emergency	6
Deaths	7
Std error of weighted % of respondents reporting CLI	8
Confirmed Cases	9
Std error of unweighted % of respondents reporting CLI	10
Mobility Changes: Parks	11
Mobility Changes: Transit	12
Mobility Changes: Retail/Recreation	13
Mobility Changes: Workplaces	14
Mobility Changes: Grocery Stores/Pharmacies	15

For Average Documented Contact Rate

Feature Name	Median Rank
Confirmed Cases	1
Cases Mean to Sum Ratio	2
Deaths	3
Number discharged from hospitals	4
Day of Week	5
Deaths Mean to Sum Ratio	6

For Hospitalization Rate

Feature Name	Median Rank
Std error of weighted % of respondents reporting CLI	1
Std error of unweighted % of respondents reporting CLI	2
Weighted % of respondents reporting CLI	3
Deaths Mean to Sum Ratio	4
Confirmed Cases	5
Cases Mean to Sum Ratio	6
Deaths	7
Day of Week	8

For Documented Recovery Rate

Feature Name	Median Rank
Deaths Mean to Sum Ratio	1
Number discharged from hospitals	2
Cases Mean to Sum Ratio	3
Confirmed Cases	4
Day of Week	5
Deaths	6

For Recovery Rate from Hospitals

Feature Name	Median Rank
Cases Mean to Sum Ratio	1
Deaths	2
Deaths Mean to Sum Ratio	3
Day of Week	4
Number discharged from hospitals	5
Confirmed Cases	6

For Death Rate from Hospitals

Feature Name	Median Rank
Mobility Changes: Residences	1
Std error of unweighted % of respondents reporting CLI	2
R_{eff}	3
Confirmed Cases	4
Cases Mean to Sum Ratio	5
State of Emergency	6
Mobility Changes: Transit	7
Std error of weighted % of respondents reporting CLI	8
Weighted % of respondents reporting CLI	9
Mobility Changes: Workplaces	10
Deaths Mean to Sum Ratio	11
Mobility Changes: Retail/Recreation	12
Deaths	13
Mobility Changes: Parks	14
Mobility Changes: Grocery Stores/Pharmacies	15

For Average Undocumented Contact Rate

Feature Name	Median Rank
Std error of unweighted % of respondents reporting CLI	1
Confirmed Cases	2
Std error of weighted % of respondents reporting CLI	3
Cases Mean to Sum Ratio	4
Deaths	5
Number of people tested	6
Weighted % of respondents reporting CLI	7
Deaths Mean to Sum Ratio	8
Day of Week	9

For Diagnosis Rate

Feature Name	Median Rank
Weighted % of respondents reporting CLI	1
Std error of weighted % of respondents reporting CLI	2
Std error of unweighted % of respondents reporting CLI	3
Confirmed Cases	4
Deaths	5
Deaths Mean to Sum Ratio	6
Cases Mean to Sum Ratio	7

For Undocumented Recovery Rate

Feature Name	Median Rank
Confirmed Cases	1
Deaths	2
Deaths Mean to Sum Ratio	3
Cases Mean to Sum Ratio	4
Day of Week	5

For Documented Death Rate

Supplementary Figure 34 | Ranking of time-varying features for encoders for Encoders in the Japan Model of January 20, 2021.

5 Ablation Studies

Table 29 shows the MAE and MAPE difference in deaths and confirmed cases with the removal of major constituents of our model. We demonstrate that partial teacher-forcing is highly important – without it, death forecasts would get much worse. This can be attributed to the rapidly-varying nature of time dynamics, and the robustness benefit of learning how to mitigate error propagation during training. Using covariates improve both death and confirmed forecasts as the model can more flexibly fit to different locations and timesteps as opposed to using constant rates, as in standard compartmental modeling. Modeling the hospitalization compartments seem to improve the death forecasts, while being less important for confirmed as expected.

Supplementary Table 29 | Ablation studies, showing the mean and 95 % confidence intervals of the MAE or MAPE gains aggregated along the prediction dates. Rows correspond to the baseline model performance, and ablation of (i) removal of using teacher forcing during training, (ii) using static rate variables for compartmental transitions (i.e. not utilizing covariates), and (iii) removal of hospitalized, ICU, and ventilator compartments, and supervision from corresponding data sources.

	Death		Confirmed	
	MAPE (%)	MAE	MAPE (%)	MAE
Baseline	5.08 ± 1.22	82.76 ± 13.00	5.85 ± 1.06	8363.75 ± 2178.59
Without teacher forcing	$+34.87 \pm 9.36$	$+951.4 \pm 238.2$	$+9.57 \pm 6.23$	$+13367.87 \pm 6226.95$
With static rates	$+1.57 \pm 2.86$	$+46.25 \pm 30.78$	$+2.17 \pm 3.66$	$+2215.74 \pm 4009.00$
Without hospitalization	$+1.00 \pm 1.06$	$+16.04 \pm 7.52$	$+0.13 \pm 0.50$	$+1002.12 \pm 1115.33$

6 Model Fairness

COVID-19 has observed to have disproportionally large impact on communities that have larger proportions of minority groups such as African-American and Hispanic subpopulations in the US [18, 19, 20, 21, 22] and it is important that our models’ forecasts are accurate across demographic groups so that and does not exacerbate these discrepancies.

When the mean performance of the model across variations in a demographic feature of interest was calculated using mean values, the counties were sorted by the feature of interest and then grouped into bins that have approximately equal numbers of geographic regions (i.e. the first bin has the quarter of the areas with the lowest proportion of the population, the second bin contains the quarter of the areas whose proportions were between the 25th and 50th quantiles, etc.). It should be noted that while this methodology compares equal numbers of counties it can lead to large disparities in the number of people in each bin which, in turn, can lead to skewed distributions in ground truth data. The demographic data for the investigation into US counties was taken from 2018 US census data and the data for the Japanese investigation used the 2015 census for sex and ethnicity [23], MHLW statistics for age [24], and the Japanese Cabinet Office statistics for median income [25].

US County Analysis

To assess the relationship of the error to each of the features of interest, the MAPE in the prospective forecasts was calculated for the 28 days forecasts and the mean values were calculated for each of the groups of counties (as described above). Bootstrapping was used to estimate the confidence intervals in each group by sampling with replacement across both the forecast date and the counties 1,000 times. The resulting mean values and confidence intervals can be seen in Tables 30 - 34.

Supplementary Table 30 | The MAPE for quartiles of US counties grouped by the race and ethnicity subgroup percentages.

Race/Ethnicity	Group Percentage	MAPE [95% CI]
Black	<0.9%	0.294 [0.240,0.364]
	0.9%-2.6%	0.237 [0.197,0.280]
	2.6%-11.0%	0.192 [0.149,0.248]
	>11.0%	0.083 [0.073,0.096]
White	<79.6%	0.102 [0.087,0.122]
	79.6%-91.4%	0.189 [0.148,0.248]
	91.4%-95.5%	0.219 [0.185,0.259]
	>95.5%	0.293 [0.239,0.360]
Hispanic	<2.4%	0.241 [0.200,0.291]
	2.4%-4.4%	0.189 [0.155,0.227]
	4.4%-10.0%	0.190 [0.154,0.235]
	>10.0%	0.181 [0.143,0.231]

Supplementary Table 31 | The MAPE for quartiles of US counties grouped by the percentage of the population that is female.

Percentage Female	MAPE [95% CI]
<49.4%	0.278 [0.222,0.344]
49.4%-50.3%	0.241 [0.195,0.308]
50.3%-51.0%	0.165 [0.140,0.195]
>51.0%	0.120 [0.100,0.145]

Supplementary Table 32 | The MAPE for quartiles of US counties grouped by median income.

Median Income	MAPE [95% CI]
<\$42.4k	0.181 [0.155,0.212]
42.5k-49.8k	0.215 [0.177,0.266]
49.9k-57.6k	0.190 [0.162,0.222]
>\$57.6k	0.216 [0.164,0.286]

An alternative methodology for quantifying the relationship between the demographic features and the models' accuracy that takes into account the continuous nature of both features is to examine the correlation of the feature of interest with the error. Kendall's Tau was used to assess

Supplementary Table 33 | The MAPE for quartiles of US counties grouped by median age.

Median Age	MAPE [95% CI]
<38.4yrs	0.134 [0.110,0.167]
38.4yrs-41.5yrs	0.179 [0.141,0.232]
41.5yrs-44.7yrs	0.214 [0.175,0.260]
>44.7yrs	0.279 [0.233,0.337]

Supplementary Table 34 | The MAPE for quartiles of US counties grouped by population density.

Population Density	MAPE [95% CI]
<6.6 ppl./sq. mi.	0.324 [0.266,0.390]
6.6-17.5 ppl./sq. mi.	0.217 [0.179,0.265]
17.5-45.9 ppl./sq. mi.	0.148 [0.127,0.173]
>46.0 ppl./sq. mi.	0.119 [0.087,0.170]

the relationship between the feature of interest and the MAPE or for each county because the MAPE values were not normally distributed and non-linear, but monotonic, relationships were also of interest [26, 27]. Significant correlations were found for all of the features other than the median income (Table 35). However, it should be noted that the most significant correlation is with the total population of the county, rather than a demographic feature, which suggests that the impact of confounding features may be significant and simply looking at changes in the mean without simultaneously accounting for the influence of other features may not be sufficient.

Supplementary Table 35 | Kendall's Tau between features of interest and the MAPE for a county.

feature	Tau	p-value	Adjusted p-value
Actual Deaths	-0.481	< 0.001	< 0.001
Total Population	-0.325	< 0.001	< 0.001
Fraction Black	-0.314	< 0.001	< 0.001
Population Density	-0.312	< 0.001	< 0.001
Fraction White	0.297	< 0.001	< 0.001
Increase in Deaths	-0.294	< 0.001	< 0.001
Fraction Female	-0.211	< 0.001	< 0.001
Median Age	0.196	< 0.001	< 0.001
Fraction Hispanic	-0.123	< 0.001	< 0.001
Median Income	-0.014	0.255	0.255

To mitigate the impact of confounding features on the correlations, we calculate the partial correlation [28] between each demographic feature of interest and the average MAPE of each county during the prospective period. To take into account the interdependency between features when considering conditional outcomes, all of the other features were included as confounding variables for partial correlation estimations. As was done above Kendall's Tau was used to quantify the remaining correlation between the MAPE and the feature of interest. We note that while this methodology helps control for the impact of confounding variables, it may also mask the impact of a single demographic variable if strong linear correlations between the factors exist.

Table 36 demonstrates that while some of the demographic features of interest were found to be statistically significant, the only feature with a correlation that had a magnitude larger than 0.1 was the population density of the county and none of the correlations with the fractions of racial or ethnic groups were found to be statistically significant at the 0.025 level.

Supplementary Table 36 | The partial correlations between features of interest and the MAPE for US counties after conditioning on the other demographic features.

feature	Tau	p-value	Adjusted p-value
Population Density	-0.152	< 0.001	< 0.001
Median Age	0.072	< 0.001	< 0.001
Median Income	-0.059	< 0.001	< 0.001
Fraction Female	-0.041	0.001	0.003
Fraction White	-0.025	0.037	0.112
Fraction Hispanic	-0.024	0.047	0.112
Fraction Black	-0.010	0.397	0.397

Japan Prefecture Analysis

As was performed for the US county-level analysis, the MAPE of the prospective 28-day forecasts of death calculated for Japan, and the aggregate statistics were calculated for each of the groups of the prefectures. The resulting mean values and confidence intervals can be seen in Tables 37 - 41.

Supplementary Table 37 | The MAPE for quartiles of Japanese prefectures grouped by the race percentages.

Race	Percentage	MAPE [95% CI]
Japanese	<98.6%	0.086 [0.066,0.109]
	98.6%-99.2%	0.169 [0.094,0.264]
	99.2%-99.5%	0.215 [0.130,0.326]
	>99.5%	0.223 [0.167,0.285]
Korean	<0.1%	0.239 [0.155,0.329]
	0.1%-0.1%	0.130 [0.083,0.180]
	0.1%-0.2%	0.172 [0.111,0.241]
	>0.2%	0.140 [0.079,0.230]
Chinese	<0.2%	0.203 [0.142,0.268]
	0.2%-0.3%	0.253 [0.154,0.372]
	0.3%-0.3%	0.137 [0.083,0.200]
	>0.4%	0.096 [0.063,0.139]

When Kendall's Tau was used to quantify the correlation between the feature of interest and the APE of each Japanese prefecture. Significant correlations at the 95% confidence level were found for the total number of actual deaths in each prefecture, the percentage of the population that was in the middle age group, the percentage of the population in the older group, the population density, and the percentage of the population that was Chinese (Table 42) after p-values were adjusted using the Holm–Bonferroni method. However if the partial correlations are computed

Supplementary Table 38 | The MAPE for quartiles of Japanese prefectures grouped by the percentage of the population that is female.

Percentage Female	MAPE [95% CI]
<50.8%	0.104 [0.065,0.156]
51.0%-51.8%	0.170 [0.088,0.277]
51.8%-52.7%	0.213 [0.125,0.308]
>52.7%	0.201 [0.148,0.264]

Supplementary Table 39 | The MAPE for quartiles of Japanese prefectures grouped by median income.

Median Income	MAPE [95% CI]
<¥257X	0.212 [0.142,0.284]
¥258X-¥285X	0.217 [0.144,0.306]
¥285X-¥301X	0.102 [0.058,0.153]
>¥303X	0.151 [0.093,0.242]

Supplementary Table 40 | The MAPE for Japanese prefectures grouped by the percentage of each age group.

Age Group	Range	MAPE [95% CI]
Younger	<11.7%	0.216 [0.134,0.323]
	11.7%-12.2%	0.128 [0.076,0.222]
	12.3%-12.7%	0.182 [0.129,0.250]
	>12.7%	0.160 [0.101,0.224]
Middle	<55.7%	0.243 [0.167,0.329]
	55.9%-57.3%	0.235 [0.156,0.331]
	57.5%-59.0%	0.109 [0.081,0.136]
	>59.3%	0.098 [0.062,0.149]
Older	<28.8%	0.100 [0.063,0.153]
	28.9%-30.4%	0.125 [0.087,0.171]
	30.6%-32.0%	0.246 [0.164,0.345]
	>32.0%	0.225 [0.137,0.326]

Supplementary Table 41 | The MAPE for quartiles of Japanese prefectures grouped by population density.

Density	MAPE [95% CI]
<179.0 ppl./sq. km	0.252 [0.187,0.317]
182.0-270.0 ppl./sq. km	0.200 [0.112,0.326]
275.0-475.0 ppl./sq. km	0.155 [0.106,0.205]
>478.0 ppl./sq. km	0.084 [0.057,0.117]

using the other demographic features as features none of the features are statistically significant (Table 43).

Supplementary Table 42 | Kendall's Tau between features of interest and the MAPE in the Japanese prefecture.

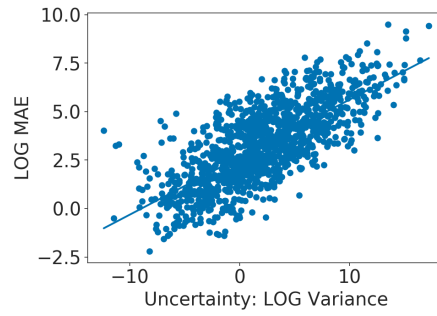
Feature	Tau	p-value	Adjusted p-value
Actual Deaths	-0.479	< 0.001	< 0.001
Middle	-0.447	< 0.001	< 0.001
Older	0.413	< 0.001	0.001
Population Density	-0.345	0.001	0.007
Chinese	-0.337	0.001	0.008
Japanese	0.297	0.004	0.026
Percentage Female	0.256	0.012	0.073
Total Population	-0.252	0.013	0.073
Median Income	-0.229	0.025	0.099
Increase in Deaths	-0.223	0.029	0.099
Korean	-0.159	0.118	0.236
Younger	-0.069	0.501	0.501

Supplementary Table 43 | The partial correlations between features of interest and the MAPE for Japanese prefectures after conditioning on the other demographic features.

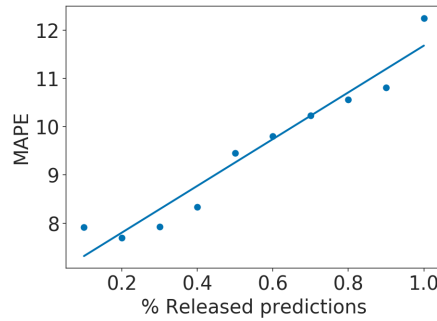
Feature	Tau	p-value	Adjusted p-value
Percentage Female	-0.244	0.017	0.149
Population Density	0.235	0.021	0.171
Korean	0.154	0.132	0.925
Japanese	0.123	0.229	1.000
Chinese	-0.074	0.466	1.000
Older	-0.063	0.538	1.000
Median Income	0.061	0.551	1.000
Middle	-0.061	0.551	1.000
Younger	-0.057	0.576	1.000

7 Uncertainty Analysis

The ability for a model to distinguish what it does and doesn't know, is an important characteristic for being able to make more reliable predictions [29, 30, 31, 32]. More formally, we can analyze this under an uncertainty quantification framework by measuring both data uncertainty and model uncertainty, where, under a fixed dataset, these are also referred to as aleatoric (data) and epistemic (model) uncertainties, respectively. Aleatoric uncertainty is irreducible uncertainty inherent in the data due to incomplete and unknown information, and is represented as a conditional distribution of the outputs conditioned on the observed inputs. Epistemic uncertainty is reducible uncertainty over the correct model, and is represented as a distribution over model functions. For our analysis of the reliability of our COVID-19 forecasts, we look into the calibration of our predicted uncertainties, specifically looking at the ability for our model uncertainty to be used for improving robustness by identifying and withholding unreliable forecasts.



Supplementary Figure 35 | Model disagreement due to model uncertainty, measured as average prediction variance across the top $k = 5$ models, versus the MAE performance, both plotted in log space. From this, we see that higher model disagreement correlates with worse metric performance. For the best fit line, $R^2 = 0.539$, $4.39x + 3.37$.



Supplementary Figure 36 | A rejection diagram showing the percentage of dates on which a prediction is made, after thresholding on model disagreement due to model uncertainty, versus the MAPE performance on those dates. From this, we can see that better average metric performance (on the days for which a forecast is released) can be achieved by withholding forecasts on days with higher model disagreement. Thus, we find the reliability of the forecasting system can be improved through model uncertainty thresholding. For the best fit line, $R^2 = 0.941$, $f(x) = 2.18x + 9.50$.

Proposed Method for Epistemic Uncertainty

For the model uncertainty part of our analysis, we look into the relationship between model disagreement and metric performance. Here, model disagreement is measured as the variance in predictions from the top- k best models, averaged over a 28-day forecast. More specifically, the procedure is as follows. For each day in the retrospective period, we start with our standard setup for training models and releasing a 28-day forecast using the best model identified during hyperparameter tuning. We then make one modification: instead of using the single best model, we instead keep the top- k best models and generate a 28-day forecast with each, where we set $k = 5$. At evaluation time, we record two values for each location: (1) the metric performance of the single best model over the 28-day forecast (MAE or MAPE), and (2) the variance in predictions across the k models for each day, averaged over the 28-day period. The variance in predictions can be interpreted as a measure of model disagreement due to model uncertainty. We repeat this for all days in the retrospective period, yielding a tuple of metric performance and

average prediction variance for each 28-day forecast release date for each location.

We then analyze this data in two ways. In Figure 35, we plot the average prediction variance versus the MAE metric performance on predicted confirmed cases, for all release dates and all prefectures in Japan. We can see that there is a positive correlation between the average prediction variance and the final metric performance. That is, higher model disagreement correlates with worse metric performance. We note though that the relationship is weaker when measured with MAPE instead of MAE, and that the correlation is noisier for some prefectures. Overall, since the prediction variance can be measured when we release a 28-day forecast, and thus before any ground truth can be observed, this correlation suggests that we could consider taking an action to mark 28-day forecasts with higher model disagreement as likely to have worse metric performance, i.e., to likely be less reliable.

To test this, we simulate the scenario of deciding, for each day, whether or not to withhold the 28-day forecast based on its associated model disagreement. We start with the data on metric performance and prediction variance from the retrospective period for Japan, as described above. For each location, we collect the set of average predicted variances for all release dates and compute ten quantiles at the $[10\%, 20\%, \dots, 90\%, 100\%]$ levels. We then decide on which forecast dates to withhold predictions by thresholding the average predicted variance based on the value at a given quantile. More specifically, for each quantile q , we threshold the set of forecast dates, and release predictions on only those dates for which the associated uncertainty value is less than or equal to the uncertainty value associated with q . This yields ten groups of dates per location representing the bottom $[10\%, 20\%, \dots, 90\%, 100\%]$, respectively, based on average predicted variance. We then compute the average metric performance for each group, and average over locations. Overall, this yields average metric performance at ten quantiles of uncertainty values, which correspond to making predictions for ten percentages of release dates.

In Figure 36, we plot this in the form of a rejection diagram [33, 34], which measures the percentage of dates on which a prediction is made (after thresholding on uncertainty) versus the average metric performance for those dates. From this, we can see that withholding forecasts on days with higher average predicted variance can lead to higher average metric performance on the remaining forecasts. We note that this overall relationship holds when measured with either MAPE or MAE. When examining single prefectures, the relationship holds when measured with MAE, but is noisier when measured with MAPE. Regardless, our conclusion is that on average, the reliability of our forecasting system can be improved through model uncertainty thresholding, at the cost of only releasing new 28-day forecasts on a lower percentage of days. Importantly, this thresholding does not rely on ground truth values, and therefore can be done in a prospective fashion as well.

8 Model applicability and case studies

We conducted a structured survey to understand how organisations were using our model forecasts. In total seven organisations responded. Their responses are summarised in Supplementary Table 44.

Supplementary Table 44 | Results of a structured survey on the applicability of the model

Organisation type	Academia (3), Government (2), Private industry (2)
Main user job role	Analyst/Scientist (3), Healthcare Professional (1), Statistician (2), Managerial (1)
Location	USA (4), Japan (3)
Predictions used	Confirmed cases (7), Death (4), Hospitalizations (4)
Model use case	ICU (3), Ventilator (2), Infected (2)
Frequency of use	Resource allocation (2), business planning (2), scenario planning (1), general understanding of COVID spread (1), confirm existing forecasts (1)
Was the model helpful	Daily (1), Weekly (1), Monthly (1)
	Yes (7)

Case studies

Estimating state and national COVID-19 testing targets

"We have an international dashboard at HGHI and Brown School of Public Health that combines a seven-day moving average of cases, hospitalizations, and deaths with forecasts from the Google-Harvard CPF. This dashboard has been used for scientific communication efforts by Dr. Thomas Tsai and Dr. Ashish Jha in the media.

"The HSPH and Brown School of Public Health team have used the Google-Harvard CPF to create estimates of state and national testing targets. To create state and national COVID-19 testing targets, we began by obtaining projected cases on January 1 from the public Google Cloud Forecasting model, and these projections formed the backbone of our entire estimate. We converted these projected cases to projected infections using a case detection rate (ratio of infections to cases) that was averaged from the Gu and CovidEstim forecast models, and then extrapolated the number of symptomatic infections and the number of contacts who would emerge from this infected population. Lastly, we incorporated testing for individuals with flu-like symptoms using CDC ILI surveillance and added allocation for surveillance testing based on the size of student, prison, nursing home, and healthcare worker populations.

"Example news coverage:

<https://www.npr.org/sections/health-shots/2020/12/22/948085513/vaccines-are-coming-but-the-u-s-still-needs-more-testing-to-stop-the-surge>

"Testing Target Dashboard:

<https://globalepidemics.org/testing-targets/january-1-2021-state-testing-targets/>

Informing state-wide social distancing measures

"We have also used the Google-Harvard CPF estimates to inform our guidance to state public health leaders. We were recently asked by Department of Public Health officials from State A to assess tradeoffs in the timing of issuing a stay-at-home order. Using the Google-Harvard CPF projections as our starting point, we were able to estimate the expected rate of exponential growth and show how a lag of a few days in implementing social distancing measures could lead to a much larger number of deaths in the ensuing weeks. After reviewing this data over the weekend,

the governor of State A implemented substantial social distancing measures on Monday morning to help curb the spread of COVID-19 in the state."

Allocating Department of Defence resources

"We used the forecasts to help the Command determine where DoD resources could be requested, and at what point requests for DoD support should be terminated based on individual state capabilities and capacity. We would compare the Google Public Forecast to others available and find the delta between them. When there were large variances, we dug into the data and local events to identify potential causes for the deviations. We used a total of 6-8 predictive models at any given time, open source research of the local areas, HHS reporting on hospital capacity, and DHS reporting on case spread/emerging hotspots.

"Google allowed us to submit specific events, such as the Sturgis Motorcycle Rally, and evaluate the potential impact on the spread of the disease through lower use of NPI, increased travel, and other factors. These "What-if Analysis" events helped us provide COVID threat assessments for multiple events that could increase disease spread. We also used the Google model to support disease threat assessments and mitigation planning for events where DoD had participation."

Making business decisions during the pandemic

"Based on the predicted number of infections by prefecture, we use the forecasts to anticipate in advance areas where infections are likely to spread, and where requests from local governments are likely to be issued. Most recently, we use the forecasts to anticipate when the Japanese Governments emergency declaration may be canceled. Based on the predicted number of confirmed cases from the infection prediction we can check whether the weekly average will meet the government's emergency declaration issuance and cancellation criteria are met.

"We want to know advance the possibility that a request from the local government will be issued, so we can better plan our sales forecasting and cost reduction. If we have the actual results of the floating population by region and day of the week, and the forecast, we think that we can take business decisions that are more effective. Our company is An SService industry with about 160 stores in more than 20 prefectures. Government requests for self-restraint from each business are being made based on the spread of infection. Where requests are issued we check the available prior information at different timings, periods, contents, and transmission methods depending on each local government. By making a prediction that a local government request will be issued in advance from the infection prediction, information can be obtained almost without omission. We were able to anticipate the request and respond according to the request in each prefecture.

"In addition, because there is an infection prediction, the management can discuss whether or not we should respond according to the request at an earlier timing in response to the request of each local government. Furthermore, we can allow each store manager to make decisions based the future spread of infection. For example, if the infection forecast increases and there is a possibility that the store may have to close or reduce it's hours (or vise versa, if infections decrease and hours can be extended), adjustments can be made to worker shift patterns in advance.

Even if we haven't decided on the above measures, we were able to convey the possibility with a certain degree of certainty, so we believe that the on-site adjustments were smooth."

References

- [1] Fda briefing document moderna covid-19 vaccine. <https://www.fda.gov/media/144434/download/>. Accessed: 2021-02-15.
- [2] David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2):281–291, 1997.
- [3] Roberto S Mariano and Francis X Diebold. Comparing predictive accuracy. *J. Bus. Econ. Stat.*, 13:253, 1995.
- [4] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.
- [5] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics*, 54(1-3):159–178, 1992.
- [6] Elizabeth A. Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1 – 21, 2010.
- [7] JM Oakes and Pamela Jo Johnson. *Propensity Score Matching for Social Epidemiology*, volume 1, pages 364–386. 05 2006.
- [8] Ariel Linden and Paul Yarnold. Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, 22:868–874, 12 2016.
- [9] Anthony Scotina and Roe Gutman. Matching algorithms for causal inference with multiple treatments. 09 2018.
- [10] Raffaele Vardavas, Aaron Strong, Jennifer Bouey, Jonathan William Welburn, Pedro Nascimento de Lima, Lawrence Baker, Keren Zhu, Michelle Priest, Lynn Hu, and Jeanne S. Ringel. *The Health and Economic Impacts of Nonpharmaceutical Interventions to Address COVID-19: A Decision Support Tool for State and Local Policymakers*. RAND Corporation, Santa Monica, CA, 2020.
- [11] Asli Demirgüç-Kunt, Michael Lokshin, and Ivan Torre. The sooner, the better: The early economic impact of non-pharmaceutical interventions during the COVID-19 pandemic. *World Bank Policy Research Working Paper*, (9257), May 2020.
- [12] Bella Nichole Kantor and Jonathan Kantor. Non-pharmaceutical interventions for pandemic COVID-19: A cross-sectional investigation of us general public beliefs, attitudes, and actions. *Frontiers in Medicine*, 7:384, 2020.

-
- [13] Holly Seale, Clare E. F. Dyer, Ikram Abdi, Kazi M. Rahman, Yanni Sun, Mohammed O. Qureshi, Alexander Dowell-Day, Jonathon Sward, and M. Saiful Islam. Improving the impact of non-pharmaceutical interventions during COVID-19: examining the factors that influence engagement and the impact on individuals. *BMC Infectious Diseases*, 20(607), 2020.
 - [14] Navid Ghaffarzadegan. Simulation-based what-if analysis for controlling the spread of COVID-19 in universities. *PLoS ONE*, 16(2), 2021.
 - [15] Pfizer and biontech conclude phase 3 study of COVID-19 vaccine candidate, meeting all primary efficacy endpoints - pfizer press release, November 2020.
 - [16] Moderna says its COVID-19 vaccine is nearly 95 percent effective, November 2020.
 - [17] J. Grauer, H. Löwen, and B. Liebchen. Strategic spatiotemporal vaccine distribution increases the survival rate in an infectious disease like COVID-19. *Nature Scientific Reports*, 10(21594), 2020.
 - [18] Clyde W. Yancy. COVID-19 and African Americans. *JAMA*, 323(19):1891–1892, 05 2020.
 - [19] Monica Webb Hooper, Anna María Nápoles, and Eliseo J. Pérez-Stable. COVID-19 and Racial/Ethnic Disparities. *JAMA*, 323(24):2466–2467, 06 2020.
 - [20] Merlin Chowkwanyun and Adolph L. Reed. Racial health disparities and COVID-19 — caution and context. *New England Journal of Medicine*, 383(3):201–203, 2020.
 - [21] Neeraj Bhala, Gwenetta Curry, Adrian R Martineau, Charles Agyemang, and Raj Bhopal. Sharpening the global focus on ethnicity and race in the time of COVID-19. *The Lancet*, 395(10238):1673–1676, 2020/07/15 2020.
 - [22] Carrie Henning-Smith, Mariana Tuttle, and Katy B Kozhimannil. Unequal distribution of COVID-19 risk among rural residents by race and ethnicity. *The Journal of rural health*, page 10.1111/jrh.12463, 05 2020.
 - [23] The Statistics Bureau of Japan. National census, 2015.
 - [24] The Statistics Bureau of Japan. The 66th Japan statistical yearbook 2017, 2015.
 - [25] Organisation for Economic Cooperation and Development. OECD.stat, 2000.
 - [26] Marie-Therese Puth, Markus Neuhäuser, and Graeme D Ruxton. Effective use of spearman’s and kendall’s correlation coefficients for association between two measured traits. *Animal Behaviour*, 102:77–84, 2015.
 - [27] Haldun Akoglu. User’s guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93, 2018.
 - [28] Ronald Aylmer Fisher et al. 035: The distribution of the partial correlation coefficient. 1924.

-
- [29] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, December 2017.
 - [30] Andrey Malinin and Mark Gales. Predictive Uncertainty Estimation via Prior Networks. In *Advances in Neural Information Processing Systems*, February 2018.
 - [31] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *arXiv:1906.02530 [cs, stat]*, June 2019.
 - [32] Michael W. Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M. Dai. Analyzing the role of model uncertainty for electronic health records. In *Proc. of the ACM Conference on Health, Inference, and Learning (ACM CHIL)*, pages 204–213, Toronto Ontario Canada, April 2020. ACM.
 - [33] Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. Benchmarking Bayesian Deep Learning with Diabetic Retinopathy Diagnosis. *arXiv:1912.10481 [stat.ML]*, 2019.
 - [34] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. *arXiv:2003.02037 [cs, stat]*, June 2020.