ARTICLE OPEN

Check for updates

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

Sercan Ö. Arık[1,9]✉, Joel Shor[2,9], Rajarishi Sinha[1,9], Jinsung Yoon[1,9], Joseph R. Ledsam[2,9], Long T. Le[1], Michael W. Dusenberry[1], Nathanael C. Yoder[1], Kris Popendorf[2], Arkady Epshteyn[1], Johan Euphrosine[2], Elli Kanal[1], Isaac Jones[1], Chun-Liang Li[1], Beth Luan[2], Joe Mckenna[1], Vikas Menon[1], Shashank Singh[1], Mimi Sun[3], Ashwin Sura Ravi[1], Leyou Zhang[1], Dario Sava[1], Kane Cunningham[1], Hiroki Kayama[2], Thomas Tsai[4], Daisuke Yoneoka[5,6], Shuhei Nomura[5,7], Hiroaki Miyata[5,8] and Tomas Pfister[1]

The COVID-19 pandemic has highlighted the global need for reliable models of disease spread. We propose an AI-augmented forecast modeling framework that provides daily predictions of the expected number of confirmed COVID-19 deaths, cases, and hospitalizations during the following 4 weeks. We present an international, prospective evaluation of our models' performance across all states and counties in the USA and prefectures in Japan. Nationally, incident mean absolute percentage error (MAPE) for predicting COVID-19 associated deaths during prospective deployment remained consistently <8% (US) and <29% (Japan), while cumulative MAPE remained <2% (US) and <10% (Japan). We show that our models perform well even during periods of considerable change in population behavior, and are robust to demographic differences across different geographic locations. We further demonstrate that our framework provides meaningful explanatory insights with the models accurately adapting to local and national policy interventions. Our framework enables counterfactual simulations, which indicate continuing Non-Pharmaceutical Interventions alongside vaccinations is essential for faster recovery from the pandemic, delaying the application of interventions has a detrimental effect, and allow exploration of the consequences of different vaccination strategies. The COVID-19 pandemic remains a global emergency. In the face of substantial challenges ahead, the approach presented here has the potential to inform critical decisions.

npj Digital Medicine (2021)4:146 ; https://doi.org/10.1038/s41746-021-00511-7

INTRODUCTION

Predicting the spread of infectious diseases is an essential component of public health management. Forecasts have contributed to resource allocation and control measures in past epi- and pandemics such as influenza[1] and Ebola[2]. Most recently, such models have shown promise during the COVID-19 pandemic[3,4] by helping ease the devastating public health and economic crisis[5–9]. However, forecasting models must overcome multiple challenges. Existing datasets contain substantial noise due to inconsistencies in reporting and the fact that many cases are asymptomatic or undocumented[10,11], and the causal impact of features within the available data is unknown. The nature of the data and the fundamental dynamics changes over time as the progression of the disease influences public policy[12] and individuals' behaviors[13] and vice versa. Beyond overcoming these, forecasting models must be explainable for decision-makers to be able to interpret the results in a meaningful way[14].

Recent work has demonstrated promising results with retrospective evaluations[3,4,15–19]. On the other hand, to understand the value of such models and their potential utility to policy decisions, a prospective evaluation is essential. Further, the utility of the forecasts needs to be rigorously validated, which is of crucial importance if such forecasts are to play a role in vaccination strategies given the wide variation in vaccine distribution, effectiveness, and uptake[20,21].

To address these challenges, we introduced an AI-augmented epidemiology framework to forecast the expected burden of COVID-19 4 weeks into the future, along with a rigorous framework for training and validation[22], and made the forecasts publicly available.

We run a prospective observational cohort study to validate the framework in the United States of America (USA) and Japan, two countries with substantial differences in healthcare systems, demographics, and the policy response to COVID-19. We demonstrate the efficacy of the framework by deriving new epidemiological findings, evaluating the predicted effect of changes in policy and behavior, and exploring settings in which the framework is being used such as hospital resource allocation and guiding state-wide social distancing policies.

RESULTS

An AI-augmented approach to epidemiology

Our framework is an extension to the Susceptible-Exposed-Infectious-Removed (SEIR) model, where a population is assigned to and may flow between compartments representing disease states[23] (Fig. 1a). Our models are optimized for the prediction of COVID-19-associated deaths and are trained on static and time-series data.

[1]Google Cloud AI, 1170 Bordeaux Dr, Sunnyvale, CA, USA. [2]Google, Japan, Shibuya, 3-Chrome-21-3, Tokyo, Japan. [3]Google Health, 1600 Amphitheatre Parkway, Mountain View, CA, USA. [4]Harvard School of Public Health, 677 Huntington Ave, Boston, MA, USA. [5]Department of Health Policy and Management, School of Medicine, Keio University, 35 Shinanomachi, Shinjuku-ku, Tokyo, Japan. [6]Division of Biostatistics and Bioinformatics, Graduate School of Public Health, St Luke's International University, 3-6-2 Tsukiji, Chuo-ku, Tokyo, Japan. [7]Department of Global Health Policy, Graduate School of Medicine, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan. [8]Department of Healthcare Quality Assessment, Graduate School of Medicine, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan. [9]These authors contributed equally: Sercan Ö. Arık, Joel Shor, Rajarishi Sinha, Jinsung Yoon, Joseph R. Ledsam. ✉email: soarik@google.com

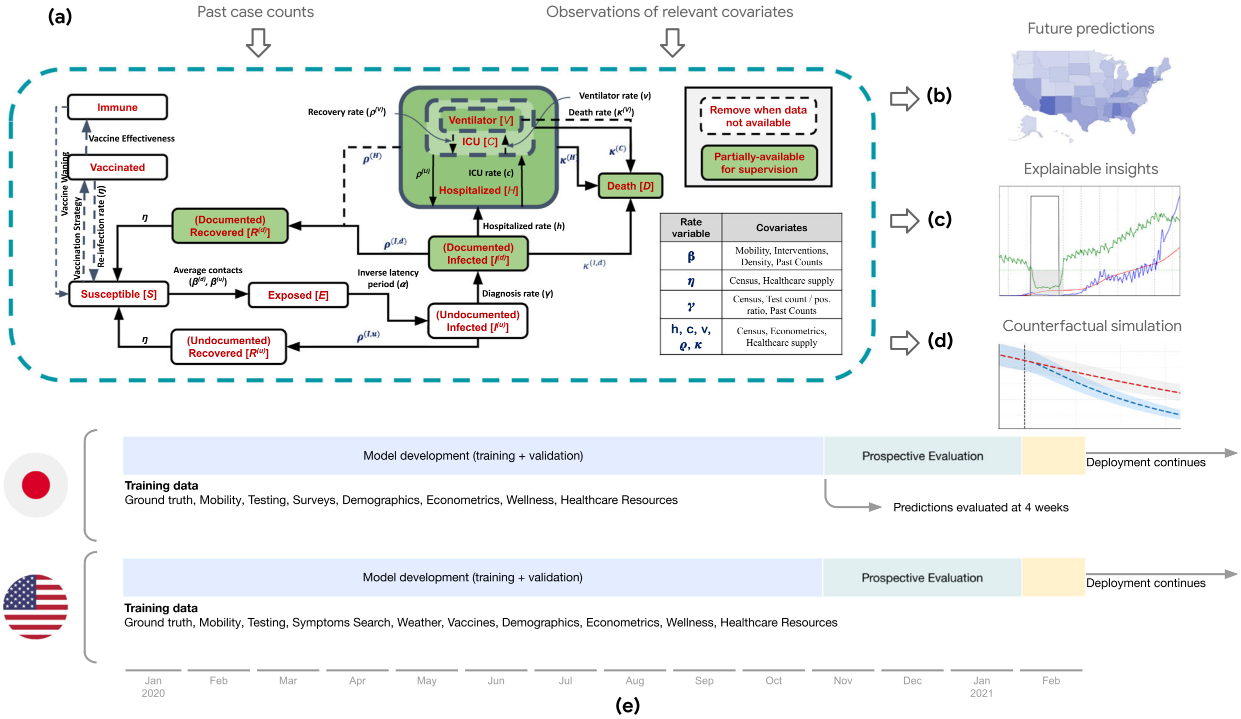
(a)
Past case counts
Observations of relevant covariates
Immune
Vaccine Effectiveness
Vaccinated
Vaccine Waning
Vaccination Strategy
Re-infection rate (η)
η
(Documented) Recovered [$R^{(d)}$]
Average contacts ($\beta^{(d)}, \beta^{(u)}$)
Susceptible [S]
Exposed [E]
Inverse latency period (α)
η
(Undocumented) Recovered [$R^{(u)}$]
$\rho^{(I,u)}$
(Undocumented) Infected [$I^{(u)}$]
Diagnosis rate (γ)
(Documented) Infected [$I^{(d)}$]
$\rho^{(I,d)}$
Hospitalized rate (h)
$\kappa^{(I,d)}$
Recovery rate ($\rho^{(V)}$)
$\rho^{(H)}$
Ventilator [V]
ICU [C]
$\rho^{(U)}$
ICU rate (c)
Hospitalized [H]
Ventilator rate (v)
Death rate ($\kappa^{(V)}$)
$\kappa^{(H)}$
$\kappa^{(C)}$
Death [D]
Remove when data not available
Partially-available for supervision
Rate variable
Covariates
β
Mobility, Interventions, Density, Past Counts
η
Census, Healthcare supply
γ
Census, Test count / pos. ratio, Past Counts
h, c, v, ϱ, κ
Census, Econometrics, Healthcare supply
(b)
Future predictions
(c)
Explainable insights
(d)
Counterfactual simulation
Model development (training + validation)
Prospective Evaluation
Deployment continues
Training data
Ground truth, Mobility, Testing, Surveys, Demographics, Econometrics, Wellness, Healthcare Resources
Predictions evaluated at 4 weeks
Model development (training + validation)
Prospective Evaluation
Deployment continues
Training data
Ground truth, Mobility, Testing, Symptoms Search, Weather, Vaccines, Demographics, Econometrics, Wellness, Healthcare Resources
Jan 2020
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan 2021
Feb
(e)

Fig. 1 Proposed framework and timeline for model development and prospective evaluation. a Our proposed AI-augmented epidemiology framework for COVID-19 forecasting is an extension to the standard Susceptible-Exposed-Infectious-Removed (SEIR) model[23,48]. We model compartments for undocumented cases explicitly as they can dominate COVID-19 spread, and introduce compartments for hospital resource usage as they are crucial to forecasts for COVID-19 healthcare planning. Learnable encoders infer the rates at which individuals move through different compartments, trained on static and time-varying public data, to model the changing disease dynamics over time and extract the predictive signals from relevant data. The models are trained daily on all available data up to the day each prediction is made (see "Methods"). **b** Public dashboard that shows generated 28-day forecasts at county- and state level for the USA. A dashboard was similarly created in Japan at the prefecture level. **c** Predictions for the effective R number and force of infection that come from the compartmental nature of the model, as well as feature importances for the rates from the variable encoder architectures. **d** Simulations of counterfactual scenarios can be used to estimate the potential impact of vaccines or policy measures. **e** Prospective evaluation of the forecasts — on each prediction date, 28-day forecasts are released publicly, and the evaluation of the accuracy is performed at the end of the 28-day horizon.

Our framework makes use of a number of key technical approaches. We use a larger number of static and time-varying features compared to prior work, powered by a systematic end-to-end learning approach with robust training mechanisms. We use a novel sequence-to-sequence modeling approach to minimize amplified growth of errors into the future due to robustness issues. Instead of depending on human-engineered functions or distribution priors to capture the impact of features, we use learnable time-varying functions to relate features to compartmental variables, which are trained in an end-to-end way. We include hospitalization compartments and demonstrate efficient training to forecast hospitalization compartments despite data sparsity. Our framework yields forecasts for a large number of locations. In our framework, a single model can provide forecasts for more than 3000 US counties with very distinct dynamics and characteristics; and indeed benefits from this transferring information across locations. Beyond point forecasts, our framework yields prediction intervals using quantile regression, and we show these are well-calibrated to capture uncertainty. We also report counterfactual simulations, and demonstrate that the counterfactual outcomes capture the relationship between the inputs and outcomes. Finally, the feature encoders in our framework can suggest the most important factors driving compartmental transitions. For additional technical details on the framework please see "Methods".

A prospective evaluation of forecasting accuracy

To evaluate our framework, we conducted a prospective observational study over 8 weeks in the USA and Japan. Predictions were made daily, each looking 4 weeks into the future (Fig. 2). We predicted the number of COVID-19 associated deaths and confirmed cases. For the USA, the available data also allowed us to predict hospitalizations, intensive care (ICU) admissions, and admissions requiring mechanical ventilation. We report both incident and cumulative mean average percentage error (MAPE) averaged across locations, as well as aggregate average percentage error (AAPE) summed over locations, in the USA and Japan (for results on absolute error, see Supplementary Discussion). Incident metrics only consider the cases that occurred during the 28-day prediction window and are helpful in the context of resource allocation. Cumulative metrics additionally take into account the total existing number of cases and can provide a broader view in the context of the pandemic thus far.

During the prospective period, across the USA as a whole, the framework achieved an AAPE of 1.4% (95% CI [1.1%, 1.6%]) for cumulative deaths and 7.1% (95% CI [5.8%, 8.5%]) for incident deaths. The framework predicted cumulative confirmed cases, incident confirmed cases, hospitalizations, intensive care unit (ICU) admissions, and admissions requiring mechanical ventilation with AAPEs of 9.20% (95% CI [8.3%, 10.2%]), 18.6% (95% CI [14.6%, 22.6%]), 59.0% [41.3%, 76.7%], 66.1% [40.2%, 92.0%], and 51.7% ([37.0%, 66.5%]), respectively. For the USA, we also provide

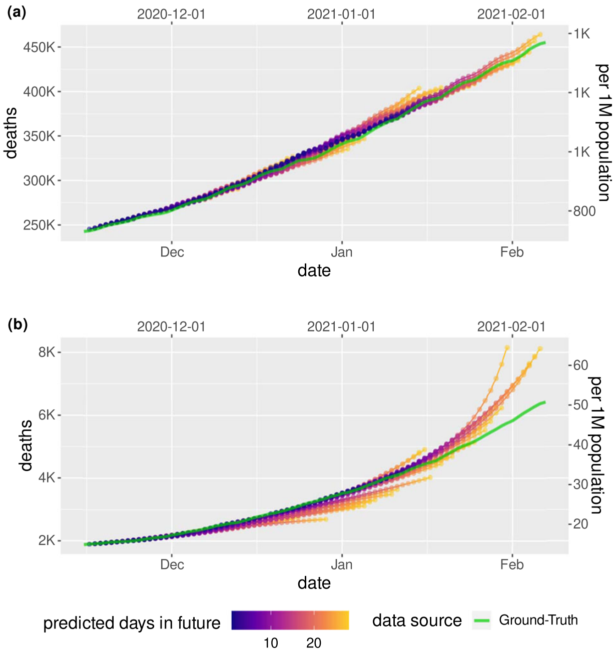

Fig. 2 Prospective forecasts for the US and Japan models. Ground truth cumulative deaths counts (cyan lines) are shown alongside the forecasts for each day. Each daily forecast contains a predicted increase in cases for each day during the prediction window of 4 weeks (shown as colored dots, where shading shifting to yellow indicates days further from the date of prediction in the forecasting horizon, up to 4 weeks). Predictions of deaths are shown for (**a**) the USA, and (**b**) Japan.

state- and county-level predictions. When evaluating at state level and averaging across all locations, the framework achieves MAPE for cumulative deaths and confirmed cases of 5.4% [5.1%, 5.6%] and 9.2% [8.2%, 10.1%], and incident deaths and confirmed cases of 24.2% [22.9%, 25.6%] and 37.8% [34.4%, 41.2%], respectively. At county level, MAPE for cumulative deaths and cumulative confirmed cases were 25.1% [23.1%, 27.0%] and 12.8% [11.5%, 14.1%], respectively. Predictions of cumulative deaths achieved an average percentage error <10% or average error <100 for 43/51 states and 2585/3006 counties, and for cumulative confirmed cases 34/51 states and 1647/3006 counties (Supplementary Tables 1 and 2).

We compare our framework with alternatives for the US (there are no public prefecture models in Japan to compare to). For statistical significance, we employ a two-sided Diebold–Mariano (DM) test (Supplementary Table 15). For cumulative deaths, the DM statistics on mean AE (MAE) are negative for all comparisons other than “Karlen-pypm”— our framework has a lower MAE in 32 out of the 33 models—and the difference with “Karlen-pypm” is statistically insignificant. Our model’s forecasts are statistically significantly different for 13/33 comparisons, and in all of these cases, our model’s MAE is lower. Using MAPE of cumulative deaths, “COVIDhub-ensemble”—a combined forecast that includes our framework’s predictions—has a negative DM statistic and a slightly lower MAPE, but this difference is not statistically significant (Supplementary Table 16). For the MAPE of incident deaths over the 4 weeks are compared, three comparisons have a negative DM statistic but none of those differences is statistically significant (Supplementary Table 17). Our model did, however, have significantly lower MAPE than five other models for incident deaths. For hospitalization predictions (Supplementary Tables 21 and 22), our model is the most accurate of the nine models, with a statistically significant difference for eight out of nine of the comparisons for MAE. For confirmed cases, the ranking of our model is lower and more variable. There are no statistically significant differences for MAE or MAPE for either cumulative or incident cases. We attribute this to the low data quality for confirmed cases, which has been widely reported[24–30], where inconsistencies in testing and reporting limit the forecastability of cases compared to hospitalizations and deaths, which have more standardized reporting. Because of this lower-quality data, we also put a lower weight on errors in confirmed cases than deaths in our multitask learning objective (see “Methods”) compared to deaths. We note that our framework can be readjusted if the desired use case places more weight on confirmed cases prediction (see “Machine-learning methods”). In addition to lower-quality data, death and hospitalizations are lagging indicators, the prediction of which may benefit more from our proposed approach than confirmed cases do.

Beyond predicting point forecasts, we also compare the quality of prediction intervals using the weighted interval score (WIS)[31,32], which is the discretized version of continuous ranking probability. Our models are in the top five and often the top model (Fig. 3, as well as Supplementary Figs. 3–11).

We also validate our model performance in Japan, reporting AAPEs for cumulative deaths and confirmed cases over the next 4-week period, 9.8% (95% CI [7.4%, 12.2%]) and 9.1% (95% CI [5.7%, 12.5%], respectively. The corresponding incident deaths and confirmed cases were 28.8% (95% CI [22.3%, 35.3%]) and 26.7% (95% CI [16.0%, 37.4%]. Data were not available on hospitalizations, ICU admissions, and admissions requiring mechanical ventilation.

The number of prefectures with an APE <10% or AE <10 for cumulative deaths and confirmed cases were 38/47 and 14/47, respectively (Supplementary Tables 3 and 4). As our models are well-calibrated, model uncertainty correlates with 28-day forecast accuracy (Fig. 4 and Supplementary Discussion). Calibrated uncertainty enables us to further improve the average prediction accuracy by flagging the uncertain predictions to be improved by human experts or other decision-making systems. For instance, we observe a 25% reduction in MAPE for cumulative deaths in Japan by only considering the most confident 50% of predictions (Fig. 4).

In addition to evaluating our framework prospectively, we also show retrospective evaluations for dates before the prospective study began. Retrospective performance was calculated by training the model using data up to a particular prediction date, and evaluating 4 weeks after the prediction date. The framework uses the most recent version that includes any corrections made to previous data and there is no leakage of data from the future dates into model training. Our comparison shows that the MAPE during the prospective period was at most 1.3% above the MAPE for the retrospective period for both cumulative deaths and cumulative confirmed cases in both the USA and Japan (Fig. 5).

We chose a 28-day prediction window to balance the timescale useful for public health decisions to be made and the rapidly changing responses to the pandemic. It also allows us to make accurate comparisons with the models at COVID-19 Forecast Hub (see “Supplementary Discussion”). However, different settings may benefit from other prediction horizons (see Supplementary Fig. 2).

COVID-19 disproportionately affects certain demographic subpopulations[33–37]. We investigate the differences in performance across locations with greater proportions of key demographic groups. While statistically significant relationships between MAPE and several demographic variables were found, only small correlations remained for most subgroups after accounting for confounding variables, suggesting that the errors are not associated with the demographic variables of race, gender, population density, or income (see the section on fairness analysis in Supplementary Note 6).

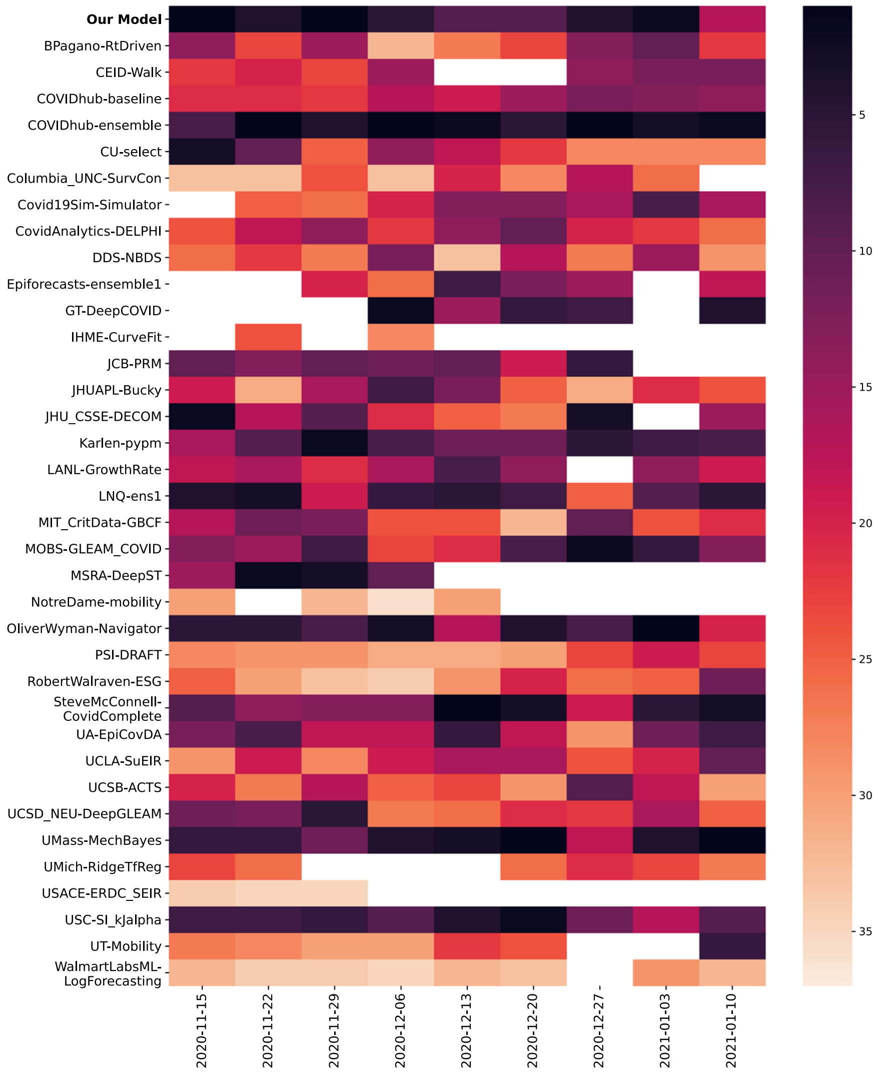

Fig. 3 Model rankings for incident death MAPE. Model rankings for incident death MAPE in the prospective evaluation period. The darker the color, the higher the ranking of the model is for the corresponding prediction date.

Ablation studies

Ablation studies demonstrate the most significant framework components that contribute to performance. Robust sequential learning with partial teacher forcing is critical for overall accuracy (yielding 34.9% ± 9.4% death and 9.6% ± 6.2% improvement in MAPE for cumulative confirmed cases). We also demonstrate the benefit of learning static and time-varying features, instead of using constant rates (yielding 1.6% ± 2.9% death and 2.2% ± 3.7% confirmed MAPE improvement). Information-sharing across locations can help generalize dynamics from other locations that have experienced similar pandemic phases in modeling confirmed cases (yielding 0.7% ± 0.8% MAPE improvement). Finally, we model hospitalization compartments, improving death forecasts (yielding 1.0% ± 1.1% MAPE improvement). See "Supplementary Discussion" for more detail.

Using the framework to understand the COVID-19 pandemic

Compartment models allow predictions of how connected compartments change over time, potentially providing information on disease dynamics including estimates for the effective reproductive number (R_{eff}), and the force of infection ($F^{(u)}$, the rate at which susceptible individuals acquire the disease). Figure 6 demonstrates this for Texas, USA and Aichi, Japan, respectively (for all other locations, see Supplementary Figs. 36–46.

We observe that nonpharmaceutical interventions (NPIs, such as mask mandates and mobility restrictions[38]) in both locations were associated with a change in R_{eff}, yielding low $F^{(u)}$ and confirmed cases. The relaxation of NPIs in Texas, and their complete removal in Aichi, were associated with cases and $F^{(u)}$ increasing. The gradual rise in the average undocumented contact rate (shown via $F^{(u)}$), results in the gradual increase in R_{eff}, which yields increasing case counts. This may also indicate that it could be beneficial to keep the NPIs in place even after $R_{eff} < 1$ while additionally observing $F^{(u)}$.

In addition, the effect of individual features on the transition rates (modeled by encoders, see "Methods") provides insights on the relative contributions of each feature (Table 1). Internet-search trends, survey results for COVID-like symptoms, and weather trends were most strongly associated with fitted contact rates. The encoder weights may also be helpful in comparing NPIs: of the seven considered for the USA, closing schools rank higher than others, suggesting its relative contribution to reducing COVID-19 spread may be greater.

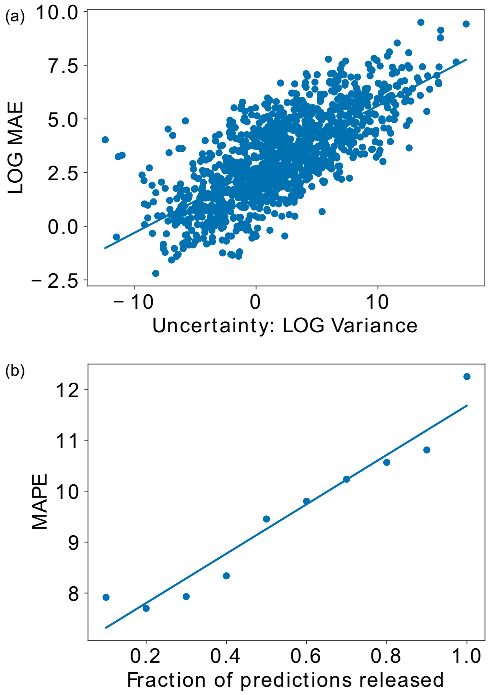

Fig. 4 **Model uncertainty. a** Model disagreement due to model uncertainty, measured as average prediction variance across the top $k = 5$ models, versus the MAE performance, both plotted in log space. From this, we see that higher model disagreement correlates with worse metric performance. For the best fit line, $R^2 = 0.539$, $4.39x + 3.37$. **b** 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, $R^2 = 0.941$, $f(x) = 2.18x + 9.50$.

Simulating the effects of interventions

Our framework can be used to predict the effect of interventions including NPIs and vaccinations. Overriding NPI features provide forecasts that simulate NPI implementation, and a "vaccinated" compartment transitioning from "susceptible" allows modeling of vaccination strategies, including dosing, effectiveness, and availability. We evaluate counterfactual accuracy by treating past NPIs as counterfactual outcomes, finding MAPE improvements when using observed features as counterfactual scenarios in all but one date tested for cumulative cases and deaths (Supplementary Table 27), as well as evaluating on simulated data ("Counterfactual" section in Supplementary Discussion). We also show uncertainty estimates for counterfactual predictions, allowing less confident estimates to be treated with appropriate caution (Supplementary Figs. 22–35).

With counterfactual analysis, consistent with the findings from feature importance, we find that school closures are associated with the highest reduction in predicted exposed counts among all NPIs. The joint application of multiple NPIs (based on the Rand levels[39]) is observed to be much more effective than each individually (Supplementary Tables 28, 29, and 37). We also find that a 7-day delay in applying all NPIs reduced predicted cases by 45% for Japan (Supplementary Table 30). Maintaining NPIs during vaccination reduces predicted cases and deaths by 9% and 9.3%. When increasing the vaccination rate to 1% of the population per day, we observe 65.5% and 16.5% reduction in susceptible and exposed counts for the US, but only a 0.8% drop in predicted cases. We do note that the overall benefit of vaccinations is more visible over longer time horizons, beyond 4 weeks. For reduction of exposed counts in the short-term, keeping particular NPIs (e.g., school closures) in place in tandem with vaccination is beneficial. The observations are also similar for Japan—keeping the State of Emergency in tandem with a high vaccination rate seems highly beneficial (Fig. 7c and Supplementary Table 35). Further results on applying and evaluating counterfactual analysis are available in Supplementary.

Use cases and the impact of our framework

Our forecasts are released publicly, and thus are available to a wide range of organizations to whom the information may aid decision-making [40]. While a robust analysis of the impact our forecasts have had is outside the scope of this paper, we conducted a structured survey of those using our forecasts to better understand how they are being used in practice. We found the forecasts, when used alongside other sources of information, were considered helpful across a broad set of areas. Uses included national resource allocation in healthcare and business settings, and implementing social distancing measures at a state-wide level. The full results of this survey, including a series of detailed case studies, are provided in Supplementary Table 56.

DISCUSSION

We present and evaluate prospectively an AI-augmented calibration procedure for compartmental models in epidemiology that forecasts at a state-, county-, and prefecture level, and provide insights relevant to current and future public health decisions. Coupled with the ability to forecast at a local level (state or county in the USA, prefecture in Japan), our framework creates the opportunity for forecasts to play a greater role in public health.

The forecasts are publicly available[40], and have been adopted alongside other information by a number of public and private organizations, alongside playing an educational role as a public reminder of the risks of COVID-19. Early case studies are positive, finding that both public and private organizations found the forecasts beneficial to a diverse range of decisions including implementing state-wide social distancing policy measures and national business decisions and healthcare resource allocation. Predictions were used alongside other available information; the forecasts are not intended to be used alone for decision-making. Despite these encouraging anecdotal reports, future quantitative studies are needed to investigate the impact of the forecasts to outcomes. Our framework also provides insight into testing resources. As our compartmental model yields the counts for undocumented and documented infected cases separately, it can suggest locations where undocumented infections are rapidly increasing, and where increasing testing may be beneficial.

Our framework provides insight into the potential consequences of public health decisions around NPIs and vaccination with counterfactual analysis. Via modifications to the proposed compartmental models, it is possible to simulate the efficacy for different vaccine regimens as new vaccines and strategies become available, helping the framework remain relevant as the pandemic evolves. This is important as our understanding of the real-world effectiveness of COVID-19 vaccines and the properties of COVID-19 variants are growing with time. The counterfactual results reinforce the risk of prematurely relaxing NPIs, the importance of prioritizing vaccinations in larger geographies, and to exercise caution even when R_{eff} is 1, until the force of infection is also reduced.

The survey conducted on organizations using the framework included academic and government organizations that had actively used this counterfactual analysis capability in their

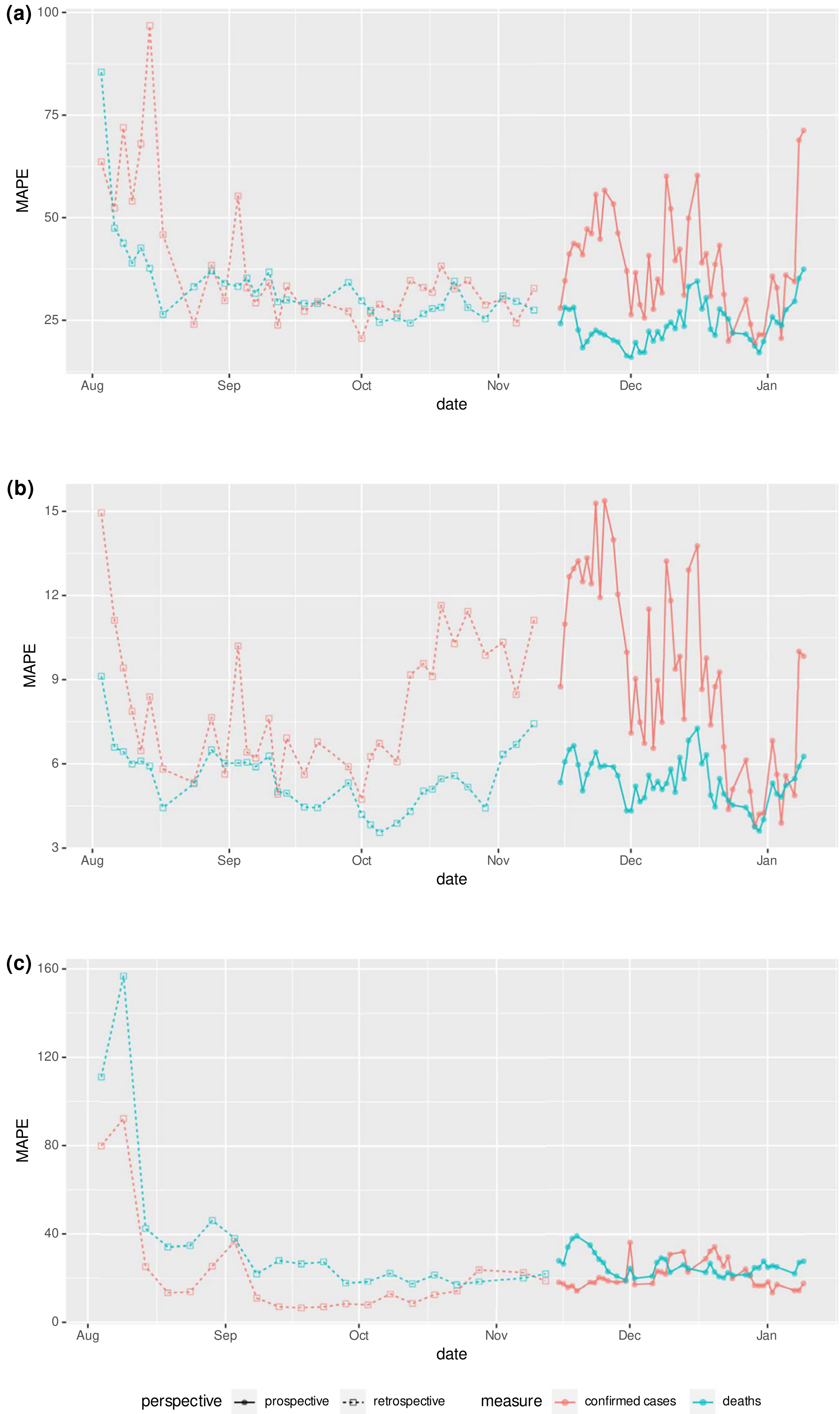

Fig. 5 Retrospective and prospective 28-day MAPE over time. Performance over time is shown for the (**a**) state-level US models using 4-week incidental predictions (**b**) state-level US models using cumulative predictions (**c**) prefecture-level Japan model using cumulative predictions. Japan's 4-week incidental deaths and cases were too low to meaningfully report. Metrics shown are the "mean absolute percentage error" for predicted deaths and predicted confirmed cases compared to ground truth. Retrospective performance during model development periods for confirmed cases (orange) and deaths (light blue) are shown alongside performance reported during the prospective study for cases (dark blue) and deaths (green).

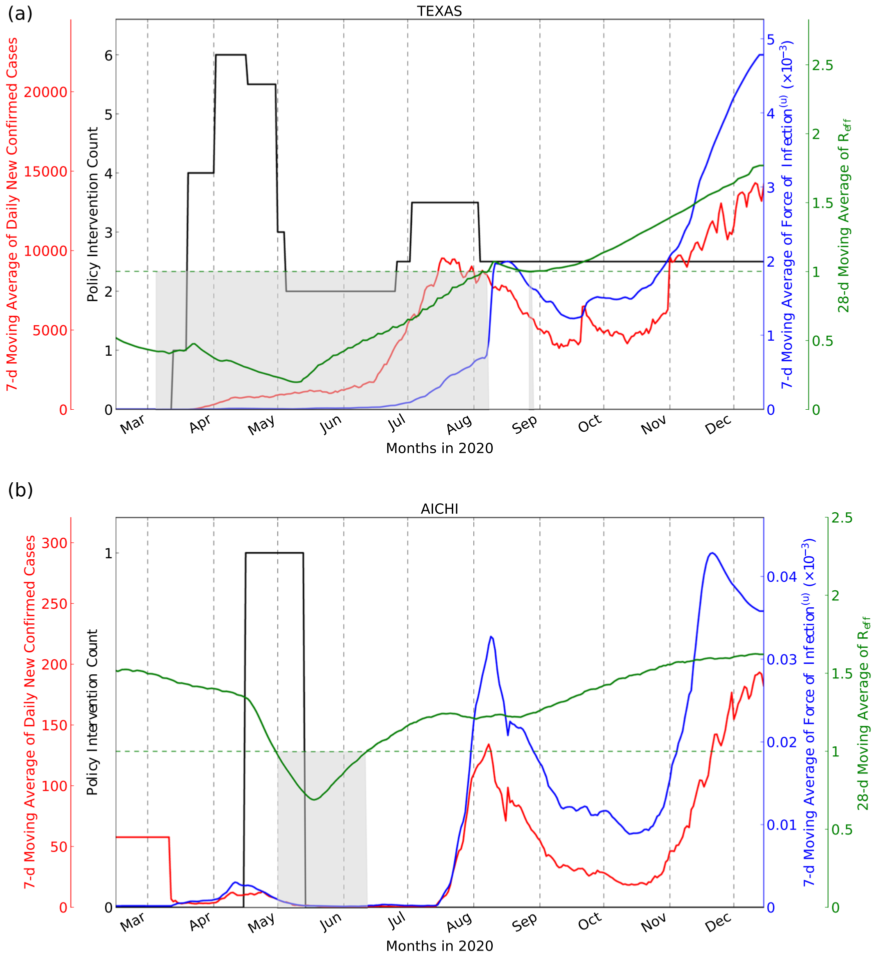

Fig. 6 Model outputs. New daily confirmed cases, number of NPIs, $F^{(u)}$ and R_{eff} for Texas, USA (**a**) and Aichi, Japan (**b**), chosen to represent a location with high and low COVID-19 associated deaths, respectively. Seven-day moving average of the daily confirmed case counts and number of Non-Pharmaceutical Interventions (NPIs) are plotted on the left y axis, and the 7-day moving average of $F^{(u)}$ (see Eq. (11)) and the 28-day moving average of R_{eff} (see Eq. (13)) are plotted on the right y axis. For $R_{eff} < 1$ (shaded gray regions below the horizontal dotted line), dynamics are tending toward the Disease-Free Equilibrium (DFE)[118]. These areas often overlap with the dates when multiple NPIs are imposed.

decision-making. Though feedback was positive, we note the uncertainty in counterfactual outcomes (which was communicated with these users) is high, and that the model outputs were used alongside other sources of information.

While the performance of the models was overall good, important variations were seen between the USA and Japan, and between different geographic locations. There are several reasons this may be the case. First, cases in Japan are skewed towards a small number of prefectures: 77.7% of Japan prefectures during prospective window had fewer than 30 deaths over the 28-day period, compared to only 0.4% for the USA. This means the model training is dominated by a small number of locations, and that evaluation using MAPE, and incident MAPE in particular, is more strongly influenced by locations with very few cases or deaths when averaging across all of Japans prefectures.

Second, there was less data to learn from due to fewer COVID-19 cases, and Japan ICU and hospitalization data were unavailable for modeling. Data quality, especially for confirmed cases, was not always consistent, including errors such as reporting delays and incorrect data[24–30]. We partially account for this with our preprocessing mechanisms (see "Methods") and by placing higher weight on confirmed deaths, which are considered to be more accurate than confirmed case counts [10,11]. However underreporting still presents challenges, as is suggested by an analysis of model R_{eff} (Supplementary Figs. 36–46).

For some states, there is a lag between peaks of R_{eff} and peaks of daily new confirmed cases, which may reflect increased testing following periods of more rapid but under-reported disease spreading. Finally, our models were optimized for predicting COVID-19-associated cumulative deaths. It is possible that performance for case prediction could be improved if the models were instead optimized for cases instead.

One potential solution to differences in performance is thresholding based on model uncertainty. Because our framework produces well-calibrated predictions, by withholding predictions when the model is uncertain, we can improve the accuracy of the remaining predictions. As each prediction provides an estimate for 4 weeks ahead, any negative impacts of withholding predictions may be relatively small.

While this publication focuses on COVID-19, our approach has value beyond the current pandemic. The underlying principles are not specific to one condition, and evidence of this is seen by the fact that performance did not substantially change during early January 2021 when new variants of SARS-CoV-2 began to emerge in both the USA and Japan. Considering future pandemics our counterfactual analysis supports existing literature on the importance of early interventions[41], and may also be useful in forward planning. Post-COVID counterfactual analyses may help better understand the relative values of different NPIs, which can be extended to novel and existing epi- and pandemics. Our results also underline the importance of making high-quality data openly

Table 1. Model feature importance.

USA model		Japan model	
Time-series features	Median rank	Time-series features	Median rank
NPI schools	1	Mobility changes: residences	1
NPI bar/restaurants	2	Std error of % of survey responders reporting CLI (unweighted)	2
Snowfall (mm)	3	Estimated R_{eff}	3
Mobility index	4	Confirmed cases	4
Cases/total tests	5	Cases mean-to-sum ratio	5
NPI non-essential business	6	State of emergency	6
Mobility samples	7	Mobility changes: transit	7
Average temperature (°C)	8	Std error of % of survey responders reporting CLI (weighted)	8
Confirmed deaths	9	% of survey responders reporting CLI (weighted)	9
Cases mean-to-sum ratio	10	Mobility changes: workplaces	10
Static features	**Median rank**	**Static features**	**Median rank**
Ratio of the population over 60	1	Average BMI of males	1
Per capita income	2	Number of H1N1 cases in 2010	2
Mean air quality index	3	Number of new ICU beds	3
Population density	4	Number of clinic beds/100 k population	4
Number of households	5	Number of doctors/100 k population	5

Top ten time-series and top five static feature importance ranks for the average undocumented contact rate in the USA and Japan models.

available[42]. For future planning, there must be coordinated efforts to make data available before it is needed.

Our work builds on a body of work in epidemiology[43–47], compartmental models[23,48–54], and machine learning[55–59]. Recent work has modeled the impact of NPIs such as travel restrictions[60] in the US[61] and Europe[62] with judiciously designed functional forms. By modeling static and time-varying features in compartmental modeling, learning their associations from data in an end-to-end way, our model improves performance while bringing explainable insights[22]. Conversely, several recent publications have attempted direct modeling from features[63–65]. In the absence of high-quality and large-scale historical data, such black-box methods underperform. The inductive bias coming from compartmental modeling has an epidemiological basis and helps the model to fit the data better. Our work differentiates from these, providing a systematic framework to ingest static and time-varying features into compartmental modeling for multi-horizon forecasting with mechanisms to inject scientific priors into the aspects that make the most sense. Our framework's performance compares favorably with alternatives (Supplementary Tables 15–22, Fig. 3, and Supplementary Figs. 10 and 11) while also providing explainability through feature encoders and counterfactual analysis, and is generalizable to higher geographic granularity and other countries. The ablation studies reported in the Supplements suggest the largest contributing factors to model performance are partial teacher forcing and the integration of learning static and time-varying features.

Our framework has several limitations. It does not differentiate between groups with different levels of risk. For vaccine modeling, differentiating the risk to priority groups (healthcare workers or the elderly population) could aid planning, but the available data do not allow this. Our models treat all locations (i.e. US states/counties and Japan prefectures) in the same way. If an application favors higher accuracy for particular locations rather than the entire country, the loss function can be tailored to overweight particular terms.

It is difficult to evaluate the accuracy of hypothetical counterfactual simulations due to the lack of ground truth. Our approach of evaluating past events and simulated data constitute only partial solutions (see "Counterfactual" section of Supplementary Discussion) as prospective evaluations of counterfactual outcomes present feasibility challenges. Rigorous experimental testing would be required to draw stronger conclusions about counterfactual accuracy. In addition to data quality, the granularity of data sources may also influence performance. One example is mobility data, where to preserve privacy only aggregated data are available. More detailed data including times of day, greater geographic granularity, or demographic factors that may influence the spread of disease could improve performance. Though we find that performance differences across locations reflect variation in case counts rather than systemic biases, the data granularity prevented evaluating subgroup performance at an individual level and biases may still be present. In addition, the uncertainty in counterfactual outcomes is high—the 95% confidence intervals for baseline and counterfactual outcomes often overlap (see Supplementary Discussion). This suggests that although the statistical significance on the directionality of the change would be high, the statistical significance on the exact amount of change would not be as high. Thus, it is important to stress that if used, the forecasts should be used alongside other information and with the support of epidemiology experts.

The COVID-19 pandemic created substantial challenges for governments, businesses, and individuals. In such an event, it is critical to inform decisions with the most accurate, up-to-date information available. We show that a generalizable, explainable AI-augmented epidemiological approach can provide accurate forecasts of the number of confirmed COVID-19 cases, deaths, and hospitalizations during the following 4 weeks, and evidence of its performance in the USA and Japan. Through our approach, we demonstrate that accurate future forecasts of case count are not only possible but are an essential and growing part of public health.

METHODS

Study design

We conduct a nationwide prospective observational study across the USA and Japan. The USA models were trained from January 22 to November 13, 2020, and the Japan model from January 15 to November 13, 2020. The study concluded on January 9, 2021 in both countries. Each daily forecast

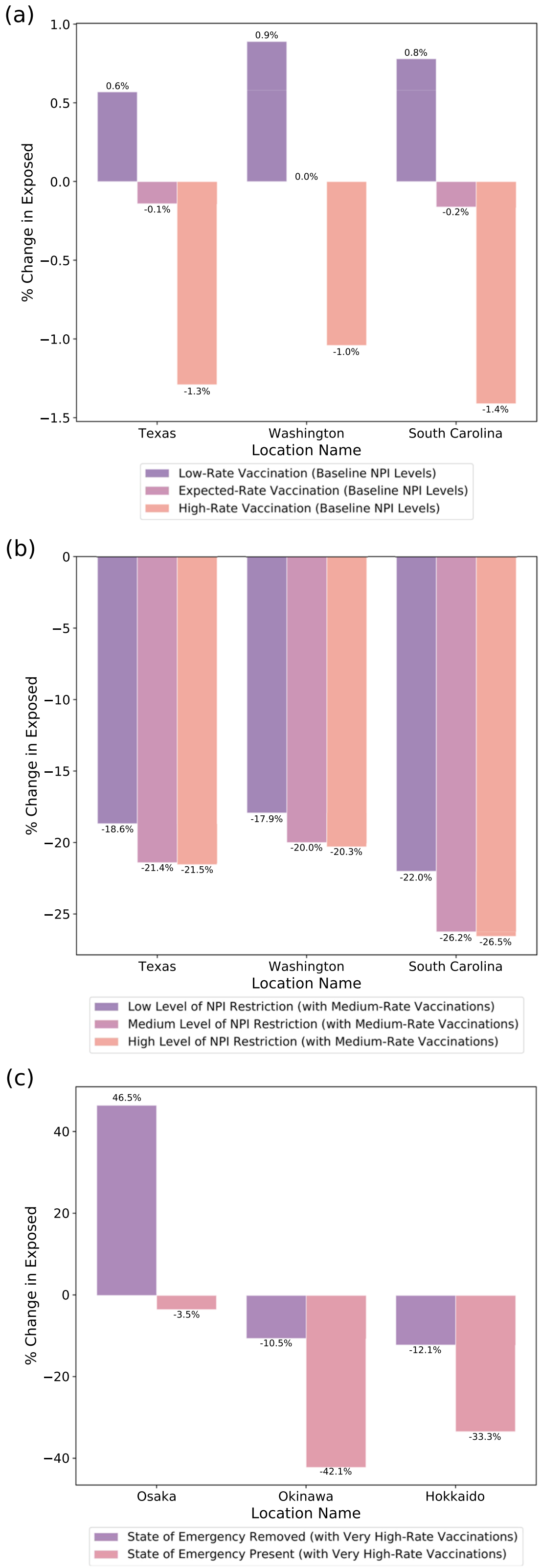

Fig. 7 Counterfactual analysis on the count of predicted exposed individuals for different vaccination rates in tandem with NPIs, for the prediction date of March 1, 2021. a As shown for the three US states, when vaccination rates (low: 0.2 % population/day, medium: 0.5% population/day, high: 1.0% population/day) are increased compared to the expected baseline, which is obtained from the past 4 weeks' trend, there is around 1% extra reduction in the predicted exposed. Here, the predicted baseline exposed individual counts are 69,700, 67,600, and 63,700 for Texas, Washington, and South Carolina, respectively. **b** For these US states, when NPI levels are increased while keeping the vaccination rate 0.5% population/day, we observe a significant reduction in the number of predicted exposed, >17% across the three states. The majority of the benefit is coming from the low-level NPI, due to the school closures being the NPI with the largest impact according to the fitted model. **c** In Japan, we show counterfactual analysis assuming a very high vaccination rate (2% population/day), and considering the cases of applying or removing the State of Emergency. Here, the baseline exposed individual counts are 5800, 3800, and 3300 for Osaka, Okinawa, and Hokkaido, respectively. Applying the state of emergency is observed to be effective in reducing the predicted exposed cases. When the State of Emergency is removed in Osaka, despite the high vaccination rate, the predicted exposed cases are observed to go up significantly. Note that in all cases, because of the uncertainty in counterfactual outcome is high—the 95% confidence intervals for baseline and counterfactual outcomes often overlap (see Supplementary Discussion). This suggests that although the statistical significance on the directionality of the change would be high, the statistical significance on the exact amount of change would not be as high. Thus, it is important to stress that if used, the forecasts should be used alongside other information and with the support of epidemiology experts. The percentage change of the exposed individual counts on March 29, 2021 against the forecasted features baseline is shown in both cases.

in this period was evaluated after 4 weeks had passed. Models were retrained daily prior to each daily forecast. All counties in the USA and all prefectures in Japan were included in the study. The entire populations of both countries were reflected in the public data; US territories were excluded. The study ran for 8 weeks, providing 56 daily forecasts to evaluate. This number was chosen based on a sample size of 43 forecasts being required to detect a 10% difference between predictions of confirmed cases and the observed values at 90% power. Of the 56 forecasts, 7 and 12 were unavailable for the USA and Japan respectively due to errors with the data sources or software bugs preventing a forecast from being produced.

Data sources and preprocessing: USA model

In this section, we describe the datasets used for our proposed framework in the US. The next section discusses Japan. The preprocessing techniques are largely the same for both countries, while the data sources are country-specific.

The ground truth data for the compartments supervise the forecasting model via training objective functions. We also use "static" (i.e. those with values that do not vary with time) and "time-varying" (i.e. those with values that vary with time) variables as inputs, to extract information from.

The progression of COVID-19 is influenced by a multitude of static variables, including relevant properties of the population, health, environmental, hospital resources, demographics, and socioeconomic indicators. Time-varying variables such as population mobility, hospital resource usage, and public policy decisions can also be important. While variables with predictive signals would be beneficial for more accurate forecasting, indiscriminately incorporating irrelevant variables may hurt performance as it may cause overfitting, if the model fits the relationships to spurious patterns that do not generalize to the future. Therefore, from multiple datasets, we choose variables that may have high predictive signals for the particular transitions in the proposed compartmental model. Those variables are used as feature inputs to the encoders which determine the transition rates. Below, we describe which features we particularly use for the USA and Japan models (also shown in Table 2).

Ground truth for compartments. For confirmed and death cases, JHU[66] is used in our work, similar to other models[45]. They obtain the raw data from the state and county health departments. Because of the rapid progression of the pandemic, past data have often been restated, or the data collection protocols have been changed. We always use the latest version of the data available prior to training or evaluation time. Ground truth data for the hospitalization compartments, including the number of people who are in ICUs or on ventilators are obtained from the COVID Tracking Project[67].

Mobility. Human mobility within a region, for work and personal reasons, may have an effect on the average contact rates[68]. We use time-varying mobility indices provided by Descartes labs at both state- and county-level resolutions[69]. Descartes labs aggregate the movement data of individual cellphone users within a region over a 24-h period. The index is equal to the ratio of the median of the distribution of distance traveled is divided by the "normal" value of the median of the distribution during the period from February 17 to March 7, 2020. These time-series features are encoded to reflect the average contact rates ($\beta^{(d)}$, $\beta^{(u)}$), at both the state- and county level of geographic resolution.

Nonpharmaceutical interventions. Public policy decisions restricting certain classes of population movement or interaction can have a beneficial effect on restricting the progression of the disease[61], at the state level of geographic resolution. The interventions are presented in six binary-valued time series, indicating when an intervention has been activated in one of six categories–school closures, restrictions on bars and restaurants, movement restrictions, mass gathering restrictions, essential businesses declaration, and emergency declaration[70]. This time-series feature is encoded into the average contact rates ($\beta^{(d)}$, $\beta^{(u)}$).

Demographics. The age of the individual may have a significant outcome on the severity of the disease and the mortality. The Kaiser Family Foundation reports the number of individuals over the age of 60 in US counties (c19hcc-info-ext-data:c19hcc_info_public.Kaiser_Health_demographics_by_Counties_States). We encode the effect of this static feature into the average contact rate ($\beta^{(d)}$, $\beta^{(u)}$), the diagnosis (γ), re-infected (η), recovery ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$) and death rates ($\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}), at both the state- and county level of geographic resolution.

Historical air quality. Historical ambient air quality in a region can have an effect on the disease spread[71]. We use the BigQuery public dataset that comes from the US Environmental Protection Agency (EPA) that documents historical air quality indices at the county level (bigquery-public-data:epa_historical_air_quality.pm10_daily_summary). This static feature is encoded into the recovery rates (η), recovery ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$) and death rates ($\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}), at both the state- and county level of geographic resolution.

Socioeconomic indicators. An individual's economic status, as well as the proximity to other individuals in a region, may have an effect on the rates of infection, hospitalization, and recovery. The proximity can be due to high population density in urban areas, or due to economic compulsions. The USA census—available from census.gov and on BigQuery Public Datasets[72]—reports state- and county-level static data on population, population density, per capita income, poverty levels, households on public assistance (bigquery-public-data:census_bureau_acs.county_2018_5yr and bigquery-public-data:census_bureau_acs.county_2018_1yr). All of these measures affect transitions into the exposed and infected compartments ($\beta^{(d)}$, $\beta^{(u)}$), as well as the recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$) and death rates ($\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}), at both the state- and county level of geographic resolution. In addition, for the state-level model, it also influences the hospitalization rate h, ICU rate c, and ventilator rate v.

Hospital resource availability. When an epidemic like COVID-19 strikes a community with such rapid progression, local hospital resources can quickly become overwhelmed[73]. To model the impact, we use the BigQuery public dataset that comes from the Center for Medicare and Medicaid Services, a federal agency within the United States Department of Health and Human Services (bigquery-public-data:cms_medicare.hospital_general_info). These static features are encoded into the diagnosis rate (γ), recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$), re-infected rate (η) and death rate ($\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}), at both the state- and county level of geographic resolution.

Symptoms search. Google provides aggregated search data related to specific disease symptoms[74] for USA states. From these symptoms, we select seven[75] as features—cough, chills, anosmia, infection, chest pain, fever, and shortness of breath. They are encoded into the diagnosis rate (γ) and the hospitalization rate h.

Table 2. Features used by the USA and Japan models.

Feature	Transition rates the feature is used for
USA model	
Per capita income	$\beta^{(d)}$, $\beta^{(u)}$, η, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Population density	$\beta^{(d)}$, $\beta^{(u)}$, η, γ, $\rho^{(I,d)}$
Households on food stamps	η, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Population	All
Number of households	$\beta^{(d)}$, $\beta^{(u)}$, η, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Population ratio above age 60	$\beta^{(d)}$, $\beta^{(u)}$, η, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Hospital rating scale	η, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Available types of hospitals	η, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Hospital patient experience rating	η, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$, $\rho^{(V)}$, h, c, v, $\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}
Air quality measures	$\beta^{(d)}$, $\beta^{(u)}$, η, $\kappa^{(I,d)}$; also for state only: h, c, v, κ^{H}, κ^{C}, κ^{V} and for county only: γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$
Mobility indices	$\beta^{(d)}$, $\beta^{(u)}$
Weather (state only)	$\beta^{(d)}$, $\beta^{(u)}$, γ, h, $\rho^{(I,d)}$, $\rho^{(I,u)}$
Google symptoms search (state only)	γ, h
Nonpharmaceutical interventions (state only)	$\beta^{(d)}$, $\beta^{(u)}$
Total tests (state only)	γ, h
Antigen/antibody tests (state only)	$\beta^{(d)}$, $\beta^{(u)}$, γ, h, $\rho^{(I,d)}$, $\rho^{(I,u)}$
Day of the week	$\beta^{(d)}$, $\beta^{(u)}$, γ, h
Confirmed per total tests	$\beta^{(d)}$, $\beta^{(u)}$, γ, h
Lagged confirmed cases	$\beta^{(d)}$, $\beta^{(u)}$, γ, h
Lagged deaths	$\beta^{(d)}$, $\beta^{(u)}$, γ, h
Japan model	
Per capita GDP	$\beta^{(d)}$, $\beta^{(u)}$, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, h, $\kappa^{(I,d)}$, κ^{H}
Population density	$\beta^{(d)}$, $\beta^{(u)}$, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, h, $\kappa^{(I,d)}$, κ^{H}
Age distribution	$\beta^{(d)}$, $\beta^{(u)}$, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, h, $\kappa^{(I,d)}$, κ^{H}
Population	All
Healthcare resources (doctors, hospital beds, clinic beds, ICU beds)	γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, h, $\kappa^{(I,d)}$, κ^{H}
Wellness (past H1N1 infection, BMI, smokers, alcohol consumption)	$\beta^{(d)}$, $\beta^{(u)}$, η, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, h, $\kappa^{(I,d)}$, κ^{H}
Google mobility indices	$\beta^{(d)}$, $\beta^{(u)}$
State of emergency	$\beta^{(d)}$, $\beta^{(u)}$
Total tests	γ
Symptoms survey results	$\beta^{(d)}$, $\beta^{(u)}$, γ, $\rho^{(I,d)}$, $\rho^{(I,u)}$
Day of week	γ, $\rho^{(I,d)}$, $\rho^{(H)}$, h, $\kappa^{(I,d)}$, κ^{H}
Confirmed mean-to-sum ratio	$\beta^{(d)}$, $\beta^{(u)}$, γ, η, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, h, $\kappa^{(I,d)}$, κ^{H}
Deaths mean-to-sum ratio	$\beta^{(d)}$, $\beta^{(u)}$, γ, η, $\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, h, $\kappa^{(I,d)}$, κ^{H}
Discharges	$\rho^{(H)}$, h, κ^{H}
Lagged confirmed cases	$\beta^{(d)}$, $\beta^{(u)}$, γ, h
Lagged deaths	$\beta^{(d)}$, $\beta^{(u)}$, γ, h

Table 3. Modeled compartments.

Compartment	Description	Compartment	Description
S	Susceptible	$R^{(u)}$	Recovered undocumented
E	Exposed	H	Hospitalized
$I^{(d)}$	Infected documented	C	In intensive care unit (ICU)
$I^{(u)}$	Infected undocumented	V	On ventilator
$R^{(d)}$	Recovered documented	D	Death
$Z^{(1)}$	First-dose vaccinated	$Z^{(1)}$	Second-dose vaccinated
Y	Immune with vaccination	L	Re-susceptible after vaccination

Weather. The Open Covid Dataset[42] provides weather features for USA states and counties and Japanese prefectures. These include daily average temperature, rainfall, and snowfall. These are encoded into the contact rates ($\beta^{(d)}$, $\beta^{(u)}$), the diagnosis rate (γ), the hospitalization rate h, and selected recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$).

Antigen and antibody test counts. Counts for antigen and antibody tests (both positive and negative outcomes) come from the Covid Tracking Project[67]. These time-series features are encoded into the contact rates ($\beta^{(d)}$, $\beta^{(u)}$), the diagnosis rate (γ), the hospitalization rate h, and selected recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$).

Day of week. The day of the week feature accounts for the cadence of data updates during the week. This feature is used for the average contact rates ($\beta^{(d)}$, $\beta^{(u)}$), the diagnosis rate (γ), and the hospitalization rate h.

Confirmed cases and deaths. Past confirmed case counts and deaths can have an effect on the current values of these quantities. We include these as time-series features. These are encoded into the average contact rates ($\beta^{(d)}$, $\beta^{(u)}$), the diagnosis rate (γ), and the hospitalization rate h.

Data sources and preprocessing: Japan model

Ground truth for compartments. We obtain the ground truth for confirmed cases, deaths, and discharges for Japanese prefectures from the Open Covid Dataset[42].

Mobility. We use six publicly available Google Mobility Reports[76,77] time series, corresponding to retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. Each time series is an index reference to a baseline value of 100 from before the pandemic. The number of unique visitors per day to places in each of the six categories is the raw measure. The raw measure is anonymized by adding Laplace noise. For each of the six measures, the reference is constructed by computing the median of the measure for each day of the week in the 5-week range from January 3, 2020 through February 6, 2020. The ratio between the raw measure and the reference is expressed as a percentage and provided as the Google mobility time series. Negative values indicate a decrease in that category of mobility and vice versa. These are encoded into the contact rates ($\beta^{(d)}$, $\beta^{(u)}$).

State of emergency. The state of emergency is a set of Covid-related restrictions[78] that are applied by the Japanese Government on a per-prefecture basis. Local- and prefecture-level authorities in Japan have wide leeway in the interpretation of the NPI[79]. We manually map the NPI to a binary-valued time series. This is encoded into the contact rates ($\beta^{(d)}$, $\beta^{(u)}$).

Symptoms survey. The Facebook Symptoms Survey dataset[80] is a dataset of survey responses regarding Covid-like illness, which could have predictive power for the COVID-19 spread and impact. We incorporate features from this dataset encoding them into the contact rates ($\beta^{(d)}$, $\beta^{(u)}$), diagnosis rate γ, and selected recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$).

Demographics. We use various prefecture-level demographic features including population, population density[81], and age distributions[82] from the 2005 census. These are encoded as continuous variables into the contact rates ($\beta^{(d)}$, $\beta^{(u)}$), diagnosis rate γ, recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$), hospitalization rate h, and selected death rates ($\kappa^{(I,d)}$, κ^{H}).

Socioeconomic indicators. We use prefecture-level per capita GDP from 2000[83] as an socioeconomic feature. It is encoded into the contact rates ($\beta^{(d)}$, $\beta^{(u)}$), diagnosis rate γ, recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$), hospitalization rate h, and selected death rates ($\kappa^{(I,d)}$, κ^{H}).

Healthcare resources. We incorporate healthcare resource features like the number of doctors, hospital, ICU[84], and clinic beds[85], both as raw and as per capita values. These features are encoded into the diagnosis rate γ, recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$), hospitalization rate h, and selected death rates ($\kappa^{(I,d)}$, κ^{H}).

Wellness. General health-related features measured before the pandemic, like BMI[86], alcohol consumption[87], past H1N1 illness[88], and smoking habits[89]. These features are encoded into the contact rates ($\beta^{(d)}$, $\beta^{(u)}$), reinfection rate η, recovery rates ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$), hospitalization rate h, and selected death rates ($\kappa^{(I,d)}$, κ^{H}).

Day of week. The day of the week feature accounts for the cadence of data updates during the week. It is encoded into the diagnosis rate γ, selected recovery rates ($\rho^{(I,d)}$, $\rho^{(H)}$), the hospitalization rate h, and selected death rates ($\kappa^{(I,d)}$, κ^{H}).

Confirmed cases and deaths. As for the USA model, the past confirmed cases and deaths can have an effect on their current values. So we include them and their derivative features (mean-to-sum ratios) into both rates.

Data sources and preprocessing: missing data

For both USA and Japan models, the data sources were provided in real time and were at risk of missing data. To address this for time-varying features, we first apply forward-filling for the future values, and then backward-filling wherever applicable. For static features, we apply median imputation. After imputation, categorical features are mapped to integer labels, and then all features are normalized to be in [0, 1], considering statistics across all locations and timesteps since the beginning of training, January 22, 2020.

Proposed compartmental model

We adapt the standard SEIR model with some major changes, as shown in Supplementary Fig. 1.

- Undocumented infected and recovered compartments: Recent studies suggest that the majority of the infected people are not detected and they dominate disease progression[15,16,51] (as the documented ones are either self-isolated or hospitalized). An undocumented infected individual is modeled as being able to spread the disease, until being documented or recovered without being undocumented.
- Hospitalized, ICU, and ventilator compartments**:** We introduce compartments for the people who are hospitalized, in the ICU, or on a ventilator, due to the practical utility to model these[73] and there are partially available observed data to be used for supervision.
- Partial immunity: To date, there is no scientific consensus on what fraction of recovered cases demonstrates immunity to future infection. Due to reports of reinfection[90], we model the rate of reinfection from recovered compartments (though our model infers low reinfection rates).
- No undocumented deaths: We assume the published COVID-19 death counts are coming from documented cases, not undocumented.
- Population invariance: We assume that the entire population is invariant,

i.e. births and non-COVID-19 deaths are negligible in comparison to the entire population.

- Vaccination: To consider the expected consequences of vaccination strategies, following[91], we introduce a new "Vaccinated" compartment, which has a transition from the "Susceptible". Approved COVID-19 vaccines have partial effectiveness[92]. In other words, only a subset of the "Vaccinated" people would actually transition into the "Immune" compartment, while some portion would become susceptible again because of the limited immunity. The two key variables for vaccination strategy—vaccine effectiveness and the number of vaccinated per day—can be adjusted for each location separately. Note that approved vaccines may be injected in one or two doses[92], and we consider multidose regimes as the number of vaccinated individuals is provided for both first and second doses, and because our framework models the partial immunity between the first and second doses.

The modeled compartments are shown in Table 3. For a compartment X, $X_i[t]$ denotes the number of individuals in that compartment at location i and time t. We assume a fixed sampling interval of 1 day. $N[t]$ denotes the total population. Figure 1 describes transition rate variables used to relate the compartments, via the equations (we omit the index i for concision):

$$\begin{aligned} S[t]-S[t-1] &= -(\beta^{(d)}I^{(d)}[t-1]+\beta^{(u)}I^{(u)}[t-1])\tfrac{S[t-1]}{N[t-1]} \\ &\quad +\eta(R^{(d)}[t-1]+R^{(u)}[t-1])-Y[t-1], \end{aligned} \tag{1}$$

$$E[t]-E[t-1] = (\beta^{(d)}I^{(d)}[t-1]+\beta^{(u)}I^{(u)}[t-1])\frac{S[t-1]}{N[t-1]}-\alpha E[t-1], \tag{2}$$

$$I^{(u)}[t]-I^{(u)}[t-1] = \alpha E[t-1]-(\rho^{(I,u)}+\gamma)I^{(u)}[t-1], \tag{3}$$

$$I^{(d)}[t]-I^{(d)}[t-1] = \gamma I^{(u)}[t-1]-(\rho^{(I,d)}+\kappa^{(I,d)}+h)I^{(d)}[t-1], \tag{4}$$

$$R^{(u)}[t]-R^{(u)}[t-1] = \rho^{(I,u)}I^{(u)}[t-1]-\eta R^{(u)}[t-1], \tag{5}$$

$$\begin{aligned} R^{(d)}[t]-R^{(d)}[t-1] &= \rho^{(I,d)}I^{(d)}[t-1]+\rho^{(H)}(H[t-1]-C[t-1]) \\ &\quad -\eta R^{(d)}[t-1], \end{aligned} \tag{6}$$

$$\begin{aligned} H[t]-H[t-1] &= hI^{(d)}[t-1]-(\kappa^{(H)}+\rho^{(H)})(H[t-1]-C[t-1]) \\ &\quad -\kappa^{(C)}(C[t-1]-V[t-1])-\kappa^{(V)}V[t-1], \end{aligned} \tag{7}$$

$$\begin{aligned} C[t]-C[t-1] &= c(H[t-1]-C[t-1])-(\kappa^{(C)}+\rho^{(C)}+v) \\ &\quad (C[t-1]-V[t-1])-\kappa^{(V)}V[t-1], \end{aligned} \tag{8}$$

$$V[t]-V[t-1] = v(C[t-1]-V[t-1])-(\kappa^{(V)}+\rho^{(V)})V[t-1], \tag{9}$$

$$\begin{aligned} D[t]-D[t-1] &= \kappa^{(V)}V[t-1]+\kappa^{(C)}(C[t-1]-V[t-1]) \\ &\quad +\kappa^{(H)}\ (H[t-1]-C[t-1])+\kappa^{(I,d)}I^{(d)}[t-1], \end{aligned} \tag{10}$$

The transition rate variables that define the relationship between the compartments are obtained with machine-learning models that input the corresponding features, as explained in the section "Machine-learning methods".

Force of infection. The force of infection is defined as the measure of the rate at which susceptible individuals become infected[93]—for undocumented infected, formulated as:

$$F^{(u)} = \beta^{(u)} * I^{(u)}/N, \tag{11}$$

and documented infected, formulated as:

$$F^{(d)} = \beta^{(d)} * I^{(d)}/N. \tag{12}$$

Effective reproductive number. Using the Next-Generation Matrix method[94] on the proposed compartmental model, the effective reproductive number can be derived as[22]:

$$R_{eff} = \frac{\beta^{(d)}\gamma+\beta^{(u)}(\rho^{(I,d)}+\kappa^{(I,d)}+h)}{(\gamma+\rho^{(I,u)})\cdot(\rho^{(I,d)}+\kappa^{(I,d)}+h)}. \tag{13}$$

Integration of vaccination. We consider two-dose vaccination strategy[95] and define the first-dose effectiveness function as:

$$\pi^{(1)}[\tau] = \min(\pi^{(1)}_{max}, \pi^{(1)}_{max}\cdot\tau/T^{(1)}_{\pi}), \tag{14}$$

and the second-dose effectiveness function as:

$$\pi^{(2)}[\tau] = \min(\pi^{(2)}_{max}, \pi^{(1)}_{max}+(\pi^{(2)}_{max}-\pi^{(1)}_{max})\cdot\tau/T^{(2)}_{\pi}), \tag{15}$$

where $\pi^{(1)}_{max}$ and $\pi^{(2)}_{max}$ are the maximum effectiveness values of the first and second vaccines, and $T^{(1)}_{\pi}$ and $T^{(2)}_{\pi}$ are the time periods defined for effectiveness ramp-up. We use $\pi^{(1)}_{max}=0.921$, $\pi^{(2)}_{max}=0.945$, $T^{(1)}_{\pi}=T^{(2)}_{\pi}=14$ days[95]. Given the cumulative counts for first-dose-vaccinated $Z^{(1)}$ (also including the s econd-dose-vaccinated) and second-dose-vaccinated $Z^{(2)}$, we obtain the count for immune with vaccination as:

$$\begin{aligned} Y[t] = &\sum_{\tau=0}^{T^{(1)}_{\pi}-1}(\pi^{(1)}[\tau]\cdot(Z^{(1)}[t-\tau]-Z^{(1)}[t-\tau-1])+\pi^{(1)}_{max}\cdot Z^{(1)}[t-T^{(1)}_{\pi}]. \\ &+\sum_{\tau=0}^{T^{(2)}_{\pi}-1}((\pi^{(2)}[\tau]-\pi^{(1)}_{max})\cdot(Z^{(2)}[t-\tau]-Z^{(2)}[t-\tau-1]) \\ &+(\pi^{(2)}-\pi^{(1)}_{max})\cdot Z^{(2)}[t-T^{(2)}_{\pi}]-L[t-1], \end{aligned} \tag{16}$$

where $L[t]$ is re-susceptible after vaccination due to the lost immunity and obtained as:

$$L[t] = L[t-1]+Y[t]/T_L, \tag{17}$$

where T_L denotes the timescale for losing immunity. We use $T_L=180$ days[95]. Note that the impact of $L[t]$ is often negligible as the forecasting horizon of our framework is much shorter.

Machine-learning methods

Time-varying modeling of variables. Instead of using static rate variables across time to model compartment transitions, there should be time-varying functions that map them from known observations. For example, if mobility decreases over time, the $S \rightarrow E$ transition should reflect that. Consequently, we propose replacing all static rate variables with learnable functions that output their value from the related static and time-varying features at each location and timestep. We note that the learnable encoding of variables still preserves the inductive bias of the compartmental modeling framework while increasing the model capacity via learnable encoders.

Interpretable encoder architecture. In addition to making accurate forecasts, it is valuable to understand how each feature affects the model. Such explanations greatly help users from the healthcare and public sector to understand the disease dynamics better, and also help model developers to ensure the model is learning appropriate dynamics via sanity checks with known scientific studies. To this end, we adopt a generalized additive model[96] for each variable v_i from Table 2 based on features $X(v_i, t)$ at different time t. The features we consider include (i) the set of static features $\mathscr{S}$, such as population density, and (ii) $\{f[t-j]\}_{f\in\mathscr{F}_i, j=1,...,k}$ the set of time-varying features $\mathscr{F}_i$ with the observation from $t-1$ to $t-k$, such as mobility. Omitting individual feature interactions and applying additive aggregation, we obtain

$$v_i[t] = v_{i,L}+(v_{i,U}-v_{i,L})\cdot\sigma\big(c+b_i+\mathbf{w}^{\top}X(v_i,t)\big), \tag{18}$$

where $v_{i,L}$ and $v_{i,U}$ are the lower and upper bounds of v_i for all t, c is the global bias, b_i is the location-dependent bias. $\mathbf{w}$ is the trainable parameter, and $\sigma()$ is the sigmoid function to limit the range to $[v_{i,L}, v_{i,U}]$, which is important to stabilize training and avoid overfitting. We use $v_{i,L}=0$ for all variables, $v_{i,U}=1$ for β, 0.2 for α, 0.001 for η, and 0.1 for others. We note that although Eq. (18) denotes a linear decomposition for $v_i[t]$ at each timestep, the overall behavior is still highly nonlinear due to the relationships between compartments.

Feature forecasting. The challenge of using Eq. (18) for future forecasting is that some time-varying features are not available for the entire forecasting horizon. Assume we have the observations of features and compartments until T, and we want to forecast from $T+1$ to $T+\tau$. To forecast $v_i[T+\tau]$, we need the time-varying features $f[T+\tau-k{:}T+\tau-1]$ for $f\in\mathscr{F}_i$, but some of them are not observed when $\tau>k$. To solve this issue, we propose to forecast $f[T+\tau-k{:}T+\tau-1]$ based on their own past observations until T, which is a standard one-dimensional time-series forecasting for a given feature f at a given location. To this end, we employ a forecasting model based on XGBoost[97] classification or regression (based on the number of categories for the feature) with time-series input features, including the lagged features of the past 7 days plus the 2 weeks ago, and mean/max in the windows of sizes of 3, 5, 7, 14, and 21 days. We note that treating feature forecasting as a univariate time-series modeling problem in this way, has limitations. There are indeed codependencies between different features (e.g., current school closure intervention affects the future value of the mobility), and also the values for the compartments (e.g., increase in the number of deaths would affect the future value of the mobility). Utilizing these has the potential to capture more information, however, given the small amount of past data, overfitting is a concern, and this way of limiting the feature forecasting model capacity and relying on implicit modeling of such dependencies, can prevent overfitting to spurious patterns from other features, and generalize better. For most features, our proposed XGBoost-based forecasting model seems to yield highly accurate forecasts, and also the impact of more accurate feature forecasting on compartmental forecasts becomes marginal beyond some point.

Information-sharing across locations. Some aspects of the disease dynamics are location-dependent while others are not. In addition, data availability varies across all L locations—there may be limited observations to learn the impact of a feature. A model able to learn both location-dependent and independent dynamics is desirable. Our encoders in Eq. (18) partially capture location-shared dynamics via shared $\mathbf{w}$ and the global bias c. To allow the model to capture remaining location-dependent dynamics, we introduce the local bias b_i. A challenge is that the model could ignore the features by encoding all information into b_i during training. This could hurt generalization as there would not be any information-sharing on how static features affect the outputs across locations. Thus, we introduce a regularization term $L_{ls} = \lambda_{ls} \Sigma_i |b_i|^2$ to encourage the model to leverage features and c for information-sharing instead of relying on b_i. Without L_{ls}, we observe that the model would use the local bias more than the encoded features, and suffers from poorer generalization.

Learning from partially available observations. Fitting would have been easy with observations for all compartments, however, we only have access to some. For instance, $I^{(d)}$ is not given in the ground truth of USA data but we instead have, Q, the total number of confirmed cases, that we use to supervise $I^{(d)} + R^{(d)} + H + D$. Note that $R^{(ud)}, I^{(ud)}, S, E$ are not given as well. Formally, we assume the availability of the observations $Y[T_s{:}\,T]$, for $Y \in \{Q, H, C, V, D, R^{(d)}\}$, and consider forecasting the next τ days, $\hat{Y}[T+1 : T + \tau]$. Note that we use the notation $S_i[T_s{:}\,T]$ to denote all timesteps between T_s (inclusive) and T (inclusive).

Fitting objective. There is no direct supervision for training encoders, while they should be learned in an end-to-end way via the aforementioned partially available observations. We propose the following objective for range $[T_s, T_e]$:

$$L_{fit}[T_s \ : \ T_e] = \sum_{Y \in \{Q,H,C,V,D,R^{(d)}\}} \boldsymbol{\lambda}_Y \sum_{t=T_s}^{T_e-\tau} \sum_{i=1}^{\tau} \frac{\mathbb{I}(Y[t+i])}{\sum_j \mathbb{I}(Y[j])} \cdot Y[j] \cdot q(t + i - T_s; z) \cdot L(Y[t+i], \hat{Y}[t+i]). \tag{19}$$

$\mathbb{I}(\cdot) \in \{0, 1\}$ indicates the availability of the Y to allow the training to focus only on available observations. $L(,)$ is the loss between the ground truth and the predicted values (e.g., ℓ_2 or quantile loss), and $\boldsymbol{\lambda}_Y$ are the important weights to balance compartments due to its robustness (e.g., D is much more robust than others). Lastly, $q(t; z) = \exp(t \cdot z)$ is a time-weighting function (when $z = 0$, there is no time weighting) to allow the fitting to favor more recent observations and z is a hyperparameter. During training, we randomly sample T_e from $[T_s, T - \tau - 1]$ and for fine-tuning, we set T_e as T.

Constraints and regularization. Given the limited dataset size, overfitting is a concern for training high-capacity encoders. In addition to limiting the model capacity with the epidemiological inductive bias, we further apply regularization to improve generalization to unseen future data. An effective regularization is constraining the effective reproduction number R_{eff} (see Eq. (13)). There is rich literature in epidemiology on R_{eff} to give us good priors on the range of the number should be. For a reproduction number $R_{eff}[t]$ at time t, we consider the regularization

$$L_{R_{eff}}[T_s : T] = \sum\nolimits_{t=T_s}^{T} \exp\big((R_{eff}[t] - R)_+\big),$$

where R is a prespecified soft upper bound. The regularization favors the model with R_{eff} in a reasonable range in addition to good absolute forecasting numbers. In our experiments, we use $R = 5$. Without this regularization term, we have observed epidemiologically unreasonable R_{eff} values (mostly in the form of sharp peaks), especially in the early days of the pandemic when the amount of training data is small. We have empirically observed that this regularization term helps smoothing such peaks, makes the training behavior a bit more robust, and eventually yields slight improvements in forecasting accuracy. Also, we integrate the prior knowledge of disease dynamics via directional penalty regularization: (1) if the mobility increases, the average contact rates ($\beta^{(d)}, \beta^{(ud)}$) will increase, (2) as the NPIs or state of emergency (SoE) introduced, the average contact rates ($\beta^{(d)}, \beta^{(ud)}$) will decrease. The directional penalty regularization is denoted as

$$L_{dir} = \sum\nolimits_{i \in \text{Mobility}} \max(-w_i, 0) + \sum\nolimits_{j \in \text{NPIs or SoE}} \max(w_j, 0),$$

Last, ignoring the perturbation of a small local window, the trend of the forecast should be usually smooth. One commonly used smoothness constraint is penalizing the first-order difference, velocity, which is defined as $v_Y[t] = (Y[t] - Y[t-k])/k$. The first-order constraint encourages $v_Y[t] \approx v_Y[t-1]$, which causes linear forecasting, and cannot capture the rapidly growing cases. Instead, we relax the smoothness to be on the second-order difference, acceleration, which is defined as $a_Y[t] = v_Y[t] - v_Y[t-1]$. The regularization is

$$L_{acc}[T_s \ : \ T] = \sum_{Y \in \{Q,D\}} \sum_{t=T_s+1}^{T} (a_Y[t] - a_Y[t-1])^2.$$

The final objective function is

$$\begin{aligned} \mathscr{L}(T_s, T) = L_{fit}[T_s : T] \ + \ \lambda_{ls} \cdot L_{ls} \ + \ \lambda_{R_{eff}} \cdot L_{R_{eff}}[T_s : T] \\ + \lambda_{dir} \cdot L_{dir} \ + \ \lambda_{acc} \cdot L_{acc}[T_s : T], \end{aligned} \tag{20}$$

where $L_{ls} = \Sigma_i |b_i|^2$.

Partial teacher forcing. The compartmental model generates the future propagated values from the current timestep. During training, we have access to the observed values for $Y \in \{Q, H, C, V, D, R^{(d)}\}$ at every timestep, which we could condition the propagated values on, commonly known as teacher forcing[59] to mitigate error propagation. At inference time, however, ground truth beyond the current timestep t is unavailable, hence the predictions should be conditioned on the future estimates. Using solely ground truth to condition propagation would create a train-test mismatch. In the same vein of past research to mix the ground truth and predicted data to condition the projections on[98], we propose partial teacher forcing, simply conditioning $(1 - \nu\mathbb{I}\{Y[t]\})Y[t] + \nu\mathbb{I}\{Y[t]\}) \hat{Y}[t]$, where $\mathbb{I}\{Y[t]\} \in \{0, 1\}$ indicates whether the ground truth $Y[t]$ exists and $\nu \in [0, 1]$. In the first stage of training, we use teacher forcing with $\nu \in [0, 1]$, which is a hyperparameter. For fine-tuning, we use $\nu = 1$ to unroll the last τ steps to mimic the real forecasting scenario.

Model fitting and selection. The training pseudocode is presented in Algorithm 1. We split the observed data into training and validation, where the validation size is τ. τ should be smaller or equal to the forecasting horizon at inference. Although having it equal minimizes the train-test mismatch, it uses more recent samples for model selection instead of training, thus, as the optimal value, we choose it to be half of the forecasting horizon. We use the training data for optimization of the trainable degrees of freedom, collectively represented as $\boldsymbol{\theta}$, while the validation data is used for early stopping and model selection. Once the model is selected, we fix the hyperparameters and run fine-tuning on joint training and validation data, to not waste valuable recent information by using it only for model selection. For optimization, we use RMSProp as it is empirically observed to yield lower losses compared to other algorithms and providing the best generalization performance. We implement Algorithm 1 in TensorFlow at state- and county levels, using ℓ_2 loss for point forecasts. We employ[99] for hyperparameter tuning (including all the loss coefficients, learning rate, and initial conditions) with the objective of optimizing for the best validation loss, with 400 trials and we use $F = 100$ fine-tuning iterations. We choose the compartment weights $\lambda^D = \lambda^Q = 0.1$, $\lambda^H = 0.01$ and $\lambda^{R^{(d)}} = \lambda^C = \lambda^V = 0.001$. We observe our results to be not highly sensitive to these hyperparameters. At county granularity, we do not have published data for C and V, so, we remove them along with their connected variables.

Quantile regression. Besides point forecasts, prediction intervals could be helpful for healthcare and public policy planners, to consider a range of possible scenarios. To obtain prediction intervals, we adapt the quantile regression method, following the state-of-the-art multi-horizon forecasting approaches for well-calibrated uncertainty[58,100]. Our framework allows the capability of modeling prediction interval forecasts, for which we replacing the L2 loss with weighted interval loss (WIS)[31] in Eq. (19) and mapping the scalar propagated values to the vector of quantile estimates. For this mapping, we use the features $Y[t]/\hat{Y}[t]$ and $\mathbb{I}\{Y[t]\}$ for $T-\tau \leq t \leq T-1$. We obtain the quantiles applying a linear kernel on these features, followed by ReLU and cumulative summation (to guarantee monotonicity of quantiles) and lastly normalization (to match the median to the input scalar point forecast from the proposed framework). In our framework, we output the α-quantile $Q_\alpha[t]$ at time t, where $\alpha \in [0.01, 0.05, 0.1, \ldots, 0.95, 0.99]$. WIS loss is a discretization of continuous ranking probability score[31].

Counterfactual analysis

Counterfactual analysis into the forecasting horizon involves replacing the forecasted values for selected NPIs, mobility features or vaccination rates with their counterfactual counterparts. Replacement or overriding happens in the forecasting horizon. For a detailed exposition, see Supplementary Discussion.

Algorithm 1 Detailed pseudocode for model training

```
Inputs: Training forecasting horizon $\tau$, compartment observations $Q_i, H_i, C_i, V_i, D_i, R_i$ from $T_s$ until $T$, the number of fine tuning iterations $F$, loss coefficients $\lambda_{R_{eff}}, \lambda_{ls}, \lambda_{dir}$ and $\lambda_{acc}$, teacher-forcing parameter $\nu$.
Initialize trainable parameters $\theta = \{\mathbf{w_i}, c, b_i\}$, set initial best validation loss ($L_{val} = \infty$) and optimal parameters ($\theta_{opt} = \theta$)
Split $\tau$ day validation $Y_i[T-\tau:T]$ for all locations $i$, where $Y \in \{Q,H,C,V,D,R^{(d)}\}$
Sample initial conditions $\hat{E}_i[0]$, $\hat{I}_i^{(d)}[0]$, $\hat{I}_i^{(u)}[0]$, $\hat{R}_i^{(d)}[0]$, $\hat{R}_i^{(u)}[0]$, $\hat{H}_i[0]$, $\hat{C}_i[0]$, $\hat{V}_i[0]$, $\hat{D}_i[0]$
while until convergence do
    Training:
    Sample random training range $T_r$ from $[T_s, T-\tau-1]$
    for $t$ in $[1, T_r]$ do
        if $t < T_r - \tau$ then
            Partial teacher forcing for all compartments $\tilde{Y}_i[t-1] = (1-\nu\mathbb{I}\{Y[t-1]\})Y[t-1] + \nu\mathbb{I}\{Y[t-1]\})\hat{Y}[t-1]$
        else
            $\tilde{Y}_i[t-1] = \hat{Y}_i[t-1]$ for all compartments
        Compute $\hat{S}_i[t], \hat{E}_i[t], \hat{I}_i^{(d)}[t], \hat{I}_i^{(u)}[t], \hat{R}_i^{(d)}[t], \hat{R}_i^{(u)}[t], \hat{H}_i[t], \hat{C}_i[t], \hat{V}_i[t], \hat{D}_i[t]$ using Eq (1) to (10) based on $\tilde{Y}_i[t-1]$
    Compute $L_{fit}[T_s : T_r]$ using $\hat{Y}_i[t]$ and $Y_i[t]$ where $Y \in \{Q,H,C,V,D,R^{(d)}\}$
    Compute regularization losses $L_{R_{eff}}, L_{ls}, L_{dir}, L_{acc}$
    Update the trainable parameters using SGD: $\theta \leftarrow \theta - \text{RMSProp}(\nabla_\theta \mathscr{L}(T_s, T_r))$
    Model selection:
    for $t$ in $[1, T]$ do
        if $t < T - \tau$ then
            Partial teacher forcing for all compartments $\tilde{Y}_i[t-1] = (1-\mathbb{I}\{Y[t-1]\})Y[t-1] + \mathbb{I}\{Y[t-1]\})\hat{Y}[t-1]$
        else
            $\tilde{Y}_i[t-1] = \hat{Y}_i[t-1]$ for all compartments
        Compute $\hat{S}_i[t], \hat{E}_i[t], \hat{I}_i^{(d)}[t], \hat{I}_i^{(u)}[t], \hat{R}_i^{(d)}[t], \hat{R}_i^{(u)}[t], \hat{H}_i[t], \hat{C}_i[t], \hat{V}_i[t], \hat{D}_i[t]$ using Eq (1) to (10) based on $\tilde{Y}_i[t-1]$
    Compute $L_{fit}[T-\tau : T]$ using $\hat{Y}_i[t]$ and $Y_i[t]$ where $Y \in \{Q,H,C,V,D,R^{(d)}\}$
    if $L_{fit}[T-\tau : T] < L_{val}$ then
        Update the optimal parameters: $\theta_{opt} = \theta$
        Set the new best validation loss as $L_{val} = L_{fit}[T-\tau : T]$
Final fine-tuning: fine-tune with joint training and validation data:
$\theta \leftarrow \theta_{opt}$
for $F$ iterations do
    for $t$ in $[1, T]$ do
        if $t < T - \tau$ then
            Partial teacher forcing for all compartments $\tilde{Y}_i[t-1] = (1-\mathbb{I}\{Y[t-1]\})Y[t-1] + \mathbb{I}\{Y[t-1]\})\hat{Y}[t-1]$
        else
            $\tilde{Y}_i[t-1] = \hat{Y}_i[t-1]$ for all compartments
        Compute $\hat{S}_i[t], \hat{E}_i[t], \hat{I}_i^{(d)}[t], \hat{I}_i^{(u)}[t], \hat{R}_i^{(d)}[t], \hat{R}_i^{(u)}[t], \hat{H}_i[t], \hat{C}_i[t], \hat{V}_i[t], \hat{D}_i[t]$ using Eq (1) to (10) based on $\tilde{Y}_i[t-1]$
    Compute $L_{fit}[T_s : T]$ and $L_{fit}[T-\tau : T]$ using $\hat{Y}_i[t]$ and $Y_i[t]$ where $Y \in \{Q,H,C,V,D,R^{(d)}\}$
    Compute regularization losses $L_{R_{eff}}, L_{ls}, L_{dir}, L_{acc}$
    Update the trainable parameters using SGD: $\theta \leftarrow \theta - \text{RMSProp}(\nabla_\theta \mathscr{L}(T_s, T))$
    if $L_{fit}[T-\tau : T] < L_{val}$ then
        Update the optimal parameters: $\theta_{opt} = \theta$
        Set the new best validation loss as $L_{val} = L_{fit}[T-\tau : T]$
Output: Return $\theta_{opt}$
```

Evaluations

Metrics. We use two kinds of metrics: the first computes metrics per geographic region (county, state, prefecture) then aggregates via averaging. The second aggregates predictions across geography then compute metrics at a country level. These two metrics allow quantification of accuracy at the location granularity of interest, as well as detection of any machine-learning biases, such as systematic underprediction or overprediction. We define per location absolute error metric as $AE_i(T,\tau) = |\hat{D}_i[T+\tau] - D_i[T+\tau]|$ and absolute percentage error metric a $APE_i(T,\tau) = 100 \cdot I\{D_i[T+\tau]>0\}(|\hat{D}_i[T+\tau]/D_i[T+\tau]| - 1)$, where $\hat{D}_i[t]$ denote the predicted variable at time t for location i, and $D_i[t]$ is the corresponding ground truth. $I\{\}$ is an indicator function that we use to eliminate 0 counts from division (this occurs for a small subset of Japanese prefectures and US counties in earlier days for the deaths). As the average absolute error metrics across all locations, we consider the average of per location error metrics: $MAE(T,\tau) = \frac{1}{L}\sum_{i=1}^{L} AE_i(T,\tau)$ and $MAPE(T,\tau) = \frac{\sum_{i=1}^{L} APE_i(T,\tau)}{\sum_{i=1}^{L} I\{D_i[T+\tau]>0\}}$. At the country level, we first aggregate counts and predictions and define aggregated absolute errors as $AAE(T,\tau) = |\sum_{i=1}^{L} \hat{D}_i[T+\tau] - \sum_{i=1}^{L} D_i[T+\tau]|$ and $AAPE(T,\tau) = |\sum_{i=1}^{L} \hat{D}_i[T+\tau]/\sum_{i=1}^{L} D_i[T+\tau] - 1|$.

Data versions for evaluations. There are significant restatements of the past observed counts in the data. For prospective evaluations, we use the data at the end of the τ day forecasting horizon. To mimic the prospective evaluations as much as possible with the retrospective evaluations, we use the reported numbers on the prediction date for training (although later we know the restated past ground truth), and the reported numbers τ days after prediction date for evaluation.

Performance comparisons. To account for the correlations between timesteps when considering the accuracy of fitting to a time series, the two-sided DM test[101] is used to compare our models' forecasts to those of other models from "covid19-forecast-hub" (https://covid19forecasthub.org/). The 4-week-ahead forecasts are compared using MAE and MAPE after they had been averaged across all the locations for the dates when both of the models produced forecasts. This is reported for cumulative and incident deaths, cumulative and incident cases, and the number of people admitted to the hospital. The P values from the tests are adjusted using the Holm–Bonferroni method[102] to account for the multiple comparisons and KPSS tests[103] are run on the differences to examine stationarity over time.

Subgroup analysis. To account for potential confounders and biases, a subset of demographic variables are chosen for further investigation. Age, sex, income, population density, and ethnicity are investigated for both the USA and Japanese models. These variables are chosen based on known biases in how COVID-19 has affected different demographics[104–108], as well as how they may affect healthcare access[109]. To investigate these relationships differences changes in the MAPE of the forecasts were compared to the demographics from each geographical region (counties for the USA and prefectures for Japan). An initial assessment is done by grouping the counties into quartiles of the demographic variable of interest and calculating the MAPE across the groups. Kendall's Tau[110] is used to quantify the relationship between the variable of interest and the MAPE for each geographical region. Because of the presence of confounding variables and potential multicollinearity between the variables of interest, partial correlation is performed using all of the other variables as features.

Uncertainty analysis. For model reliability when used by human experts, we also investigate using epistemic model uncertainty as a confidence metric of forecasts. Well-calibrated estimates of uncertainty are important for being able to make more reliable predictions[111–115]. To this extent, we investigate the relationship between epistemic model uncertainty and the accuracy of forecasts by simulating the scenario of deciding whether or not to withhold each day's 28-day forecast based on model disagreement, with the goal of demonstrating an optional feature that could allow the system to identify and withhold the predictions that are likely to be the most erroneous.

For each day in the retrospective period, we train an ensemble of $k = 5$ models and use their disagreement as a metric of uncertainty[111]. Producing 28-day forecasts from each model, we consider two values for each location: (1) the metric performance of the single best model over the 28-day forecast (in MAE or MAPE for cumulative values), and (2) the variance in predictions across the k models for each day, averaged over the 28-day period. In Fig. 4, 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 higher model disagreement correlates with worse metric performance.

For each location, we then collect the set of average predicted variances for all release dates and compute ten quantiles at the [10%, 20%, …, 90%, 100%] levels. We then decide on which forecast dates to withhold predictions by thresholding the average predicted variance based on the value at each quantile, yielding ten groups of release dates per location. We average the metric performance across the dates in each group and average over locations. This yields average performance at ten quantiles of uncertainty values, which we visualize in the form of a rejection diagram[116,117] in Fig. 4.

From this, we can see that withholding forecasts on days with higher average prediction variance can lead to better average metric performance over the remaining forecasts.

Overall, we find that on average, the reliability of our framework could be improved through the proposed method of model uncertainty quantification by allowing the system to identify and withhold predictions that are likely to be the most erroneous at prediction time without access to ground truth labels (additional discussion in Supplementary Note: Uncertainty Analysis). Importantly, this is an optional feature, with prediction withholding being just one example of an action that the system could take upon identifying an unreliable prediction. Other actions could include bringing an expert human into the loop to further analyze the predictions for that day or simply adding a note to the released predictions to indicate that they may be less reliable.

Uncertainty in counterfactual predictions. As with the forecast uncertainty, the prediction of counterfactuals is subject to disagreement between independently trained models. We describe the process to obtain the uncertainty in the counterfactuals and some observations in the Supplementary Discussion.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data used for the training, validation, and test sets are publicly available. All data were collected entirely from openly available sources. The dashboard showing our forecasts can be accessed from https://g.co/covidforecast. The following websites can be used to access the time-series data used in this study for the US model: https://github.com/CSSEGISandData/COVID-19 (state and county confirmed cases, deaths); https://github.com/descarteslabs/DL-COVID-19 (state and county mobility, mobility index and mobility samples); https://covidtracking.com/ (state hospitalized currently, state hospitalized cumulative counts, state hospitalized increase, state in ICU currently, state in ICU cumulative, state on ventilator currently, state total test results); https://coronadatascraper.com/#home (county hospitalized cumulative counts); http://goo.gle/covid19symptomdataset (state and county symptoms search tracker); https://dsd.c19hcc.org/ (state NPI information). The static data for the US can be found at https://data.cms.gov/provider-data/dataset/xubh-q36u (state and county hospital resources); https://www.census.gov/programs-surveys/acs (state and county population and demographics data); https://www.epa.gov/ (state and county air quality data). For the Japan model time-series data, the following datasets were used: https://cloud.google.com/blog/products/data-analytics/publicly-available-covid-19-data-for-analytics (confirmed cases and deaths); https://covidmap.umd.edu/api.html (symptoms data); https://github.com/google-research/open-covid-19-data/blob/master/data/exports/google_mobility_reports/Regions/2020_JP_Region_Mobility_Report.csv (mobility data); https://crisis.ecmonet.jp/ (ventilator use); https://github.com/graphy-covidjp/graphy-covidjp.github.io (vaccine data). The static data for Japan can be found at https://stats-japan.com/t/kiji/13400 (demographics); https://www.mhlw.go.jp/english/database/db-hh/2-2.html (healthcare resources and wellness data); https://www.mhlw.go.jp/toukei/list/20-21.html (living conditions); https://stats.oecd.org/ (socioeconomic indicators); http://idsc.nih.go.jp/idwr/sokuho/index.html (H1N1 data); https://www.mhlw.go.jp/bunya/kenkou/eiyou/h24-houkoku.html (nutrition data); https://japan.kantei.go.jp/ongoingtopics/index.html (state of emergency data).

CODE AVAILABILITY

We make use of several open-source libraries to conduct our experiments, namely the machine-learning framework TensorFlow (https://github.com/tensorflow/tensorflow). While the entire system relies on proprietary libraries and we are unable to publicly release the entire codebase, we make available code for key pieces of the model details, training, and evaluation at https://github.com/google-research/google-research/tree/master/covid_epidemiology.

Received: 9 June 2021; Accepted: 25 August 2021;
Published online: 08 October 2021

REFERENCES

1. Shaman, J. & Karspeck, A. Forecasting seasonal outbreaks of influenza. *Proc. Natl Acad. Sci. USA* **109**, 20425–20430 (2012).
2. Wendlandt, M., Colabella, J. M., Krishnamurthy, A. & Cobb, L. Mathematical modelling of the west African ebola virus epidemic. *URSCA Proc.* **3**, (2017).
3. Ray, E. L. et al. Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U. S. Preprint at https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1 (2020).
4. Reiner, R. C. et al. Modeling COVID-19 scenarios for the united states. *Nat. Med.* **27**, 94–105 (2020).
5. Bedford, J. et al. COVID-19: towards controlling of a pandemic. *Lancet* **395**, 1015–1018 (2020).
6. Sinclair, A. J. & Abdelhafiz, A. H. Age, frailty and diabetes—triple jeopardy for vulnerability to COVID-19 infection. *EClinicalMedicine* **22**, 100343–100343 (2020).
7. Ji, Y., Ma, Z., Peppelenbosch, M. P. & Pan, Q. Potential association between COVID-19 mortality and health-care resource availability. *Lancet Global Health* **8**, e480 (2020).
8. Bonaccorsi, G. et al. Economic and social consequences of human mobility restrictions under COVID-19. *Proc. Natl Acad. Sci. USA* **117**, 15530–15535 (2020).
9. Borio, C. The COVID-19 economic crisis: dangerously unique. *Business Econ.* **55**, 181–190 (2020).
10. Pearce, N., Vandenbroucke, J. P., VanderWeele, T. J. & Greenland, S. Accurate statistics on COVID-19 are essential for policy guidance and decisions. *Am. J. Public Health* **110**, 949–951 (2020).
11. Fenton, N., Hitman, G. A., Neil, M., Osman, M. & McLachlan, S. Causal explanations, error rates, and human judgment biases missing from the COVID-19 narrative and statistics. Preprint at https://psyarxiv.com/p39a4/ (2020).
12. Haug, N. et al. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat. Hum. Behav.* **4**, 1303–1312 (2020).
13. Amuedo-Dorantes, C., Kaushal, N. & Muchow, A. N. Is the cure worse than the disease? County-level evidence from the COVID-19 pandemic in the united states. NBER Working Papers 27759, (National Bureau of Economic Research, Inc., 2020).
14. Ahmad, M. A., Eckert, C. & Teredesai, A. Interpretable machine learning in healthcare. in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.* 559–560 (Association for Computing Machinery, 2018). https://doi.org/10.1145/3233547.3233667.
15. Long, C., Fu, X. M. & Fu, Z. F. Global analysis of daily new COVID-19 cases reveals many static-phase countries including the United States potentially with unstoppable epidemic. *World J Clin Cases* **8**, 4431–4442. https://doi.org/10.12998/wjcc.v8.i19.4431 (2020).
16. Long, Y.-S. et al. Quantitative assessment of the role of undocumented infection in the 2019 novel coronavirus (COVID-19) pandemic. Preprint at https://arxiv.org/abs/2003.12028 (2020).
17. Chang, S. et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* **589**, 82–87 (2020).
18. Rodríguez, A. et al. DeepCOVID: An Operational Deep Learning-driven Framework for Explainable Real-time COVID-19 Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 15393–15400 (2021).
19. Subramanian, R., He, Q. & Pascual, M. Quantifying asymptomatic infection and transmission of covid-19 in new york city using observed cases, serology, and testing capacity. *Proc. Natl Acad. Sci.* **118**, e2019716118 (2021).
20. Johnson & johnson announces single-shot Janssen COVID-19 vaccine candidate met primary endpoints in interim analysis of its phase 3 ensemble trial. https://www.jnj.com/johnson-johnson-announces-single-shot-janssen-COVID-19-vaccine-candidate-met-primary-endpoints-in-interim-analysis-of-its-phase-3-ensemble-trial (2021).
21. Japanese association of infectious diseases: recommendations for COVID-19 vaccine. https://www.kansensho.or.jp/modules/guidelines/index.php?content_id=43 (2021).
22. Arik, S. O. et al. Interpretable sequence learning for COVID-19 forecasting. *Proceedings of the 34th Conference on Neural Information Processing Systems*, (NeurIPS, 2020).
23. Blackwood, J. C. & Childs, L. M. An introduction to compartmental modeling for the budding infectious disease modeler. *Lett. Biomathematics* **5**, 195–221 (2018).
24. Noh, J. & Danuser, G. Estimation of the fraction of covid-19 infected people in u. s. states and countries worldwide. *PLoS ONE* **16**, 1–10 (2021).
25. Anand, S. et al. Prevalence of sars-cov-2 antibodies in a large nationwide sample of patients on dialysis in the USA: a cross-sectional study. *Lancet* **396**, 1335–1344 (2020).
26. Lau, H. et al. Evaluating the massive underreporting and undertesting of covid-19 cases in multiple global epicenters. *Pulmonology* **27**, 110–115 (2021).
27. Wu, S. L. et al. Substantial underestimation of sars-cov-2 infection in the united states. *Nat. Commun.* **11**, 4507 (2020).
28. Bendavid, E. et al. COVID-19 antibody seroprevalence in Santa Clara County, California. *Int. J. Epidemiol.* **50**, 410–419 (2021).
29. Wang, Y.-Y. et al. Updating the diagnostic criteria of covid-19 "suspected case" and "confirmed case" is necessary. *Military Med. Res.* **7**, 17 (2020).
30. Region data notes. https://coronavirus.jhu.edu/region-data-notes.
31. Bracher, J., Ray, E. L., Gneiting, T. & Reich, N. G. Evaluating epidemic forecasts in an interval format. Bracher, J., Ray, E.L., Gneiting, T. & Reich N.G. Evaluating epidemic forecasts in an interval format. *PLoS Comput Biol* **17**, e1008618 (2021). https://doi.org/10.1371/journal.pcbi.1008618. Preprint at https://arxiv.org/abs/2005.12881 (2020).
32. Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
33. Yancy, C. W. COVID-19 and African Americans. *J. Am. Med. Assoc.* **323**, 1891–1892 (2020).
34. Webb Hooper, M., Nápoles, A. M. & Pérez-Stable, E. J. COVID-19 and racial/ethnic disparities. *J. Am. Med. Assoc.* **323**, 2466–2467 (2020).
35. Chowkwanyun, M. & Reed, A. L. Racial health disparities and COVID-19—caution and context. *New Engl. J. Med.* **383**, 201–203 (2020).
36. Bhala, N., Curry, G., Martineau, A. R., Agyemang, C. & Bhopal, R. Sharpening the global focus on ethnicity and race in the time of COVID-19. *Lancet* **395**, 1673–1676 (2020).
37. Henning-Smith, C., Tuttle, M. & Kozhimannil, K. B. Unequal distribution of COVID-19 risk among rural residents by race and ethnicity. *J. Rural Health* https://doi.org/10.1111/jrh.12463 (2020).
38. Non-pharmaceutical interventions. https://www.cdc.gov/nonpharmaceutical-interventions/index.html (2020).
39. Vardavas, R. et al. *The Health and Economic Impacts of Nonpharmaceutical Interventions to Address COVID-19: A Decision Support Tool for State and Local Policymakers*. (RAND Corporation, 2020).
40. COVID-19 public forecasts. https://pantheon.corp.google.com/marketplace/product/bigquery-public-datasets/covid19-public-forecasts (2021).
41. Demirguc-Kunt, A., Lokshin, M. & Torre, I. The sooner, the better: The early economic impact of non-pharmaceutical interventions during the COVID-19 pandemic. World Bank Policy Research Working Paper (2020).
42. Wahltinez, O. et al. COVID-19 open-data: curating a fine-grained, global-scale data repository for sars-cov-2 (2020). https://goo.gle/COVID-19-open-data (2020).
43. Capasso, V. In *Proceedings of the Second European Symposium on Mathematics in Industry* (ed. Neunzert, H.) Vol. 3, 181–194 (Springer, 1988).
44. Hunter, E., Namee, B. M. & Kelleher, J. An open-data-driven agent-based model to simulate infectious disease outbreaks. *PLoS ONE* **13**, e0208775 (2018).
45. Murray, C. J. Forecasting COVID-19 impact on hospital bed-days, icu-days, ventilator-days and deaths by us state in the next 4 months. Preprint at https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1 (2020).
46. GrowthRate model from Los Alamos National Laboratory.https://COVID-19.bsvgateway.org/#link%20to%20forecasting%20site (2020).
47. White, S. H., del Rey, A. M. & Sánchez, G. R. Modeling epidemics using cellular automata. *Appl. Mathematics Comput.* **186**, 193–202 (2006).
48. Smith, D. & Moore, L. The SIR Model for Spread of Disease - The Differential Equation Model. *Convergence* (December 2004).
49. Kermack, W. O. & McKendrick, A. G. Contributions to the mathematical theory of epidemics-i. *Proc. Royal Soc.* **115A**, 700–721 (1927).
50. Grand Rounds. COVID-19 forecasting: fit to a curve or model the disease in real-time? https://grandrounds.com/blog/COVID-19-forecasting-fit-to-a-curve-or-model-the-disease-in-real-time/. (2020).
51. Li, R. et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science* **368**, 489–493 (2020).
52. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, 2009).

53. Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D. & Del Valle, S. Y. Forecasting seasonal influenza with a state-space SIR model. *Ann. Appl. Stat.* **11**, 202 (2017).
54. Bretó, C., He, D., Ionides, E. L. & King, A. A. Time series analysis via mechanistic models. *Ann. Appl. Stat.* **3**, 319–348 (2009).
55. Greydanus, S., Dzamba, M. & Yosinski, J. Hamiltonian Neural Networks. *Proceedings of the 33rd Conference on Neural Information Processing Systems*, (NeurIPS, 2019).
56. Cranmer, M. et al. Lagrangian Neural Networks. *Workshop on Deep Differential Equations at the 8th International Conference on Learning Representations*, ICLR (2020).
57. Lutter, M., Ritter, C. & Peters, J. Deep lagrangian networks: using physics as model prior for deep learning. *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. (2019).
58. Lim, B., Arik, S. O., Loeff, N. & Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, **37**, 1748–1764 (2021).
59. Williams, R. J. & Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1**, 270–280 (1989).
60. Chinazzi, M. et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).
61. Pei, S., Kandula, S. & Shaman, J. Differential effects of intervention timing on COVID-19 spread in the united states. *Sci. Adv.* **6**, eabd6370 (2020).
62. Flaxman, S. et al. Report 13—estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. (2020).
63. Venna, S. R. et al. A novel data-driven model for real-time influenza forecasting. *IEEE Access* **7**, 7691–7701 (2019).
64. Yang, Z. et al. Modified seir and ai prediction of the epidemics trend of COVID-19 in china under public health interventions. *J. Thoracic Dis.* **12**, 165–174 (2020).
65. Wang, L., Chen, J. & Marathe, M. Tdefsi: theory guided deep learning based epidemic forecasting with synthetic information. *ACM Transact. Spatial Algorithms Syst. (TSAS)* **6.3**, 1–39 (2020)..
66. Ensheng Dong, H. & Du Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
67. Covid-Tracking. The covid tracking project. https://covidtracking.com/ (2020).
68. Bryant, P. & Elofsson, A. Estimating the impact of mobility patterns on COVID-19 infection rates in 11 European countries. Preprint at https://www.medrxiv.org/content/10.1101/2020.04.13.20063644v2 (2020).
69. Warren, M. S. & Skillman, S. W. Mobility changes in response to COVID-19. Preprint at https://arxiv.org/abs/2003.14228 (2020).
70. Realtime tracking of state-wide npi implementations. https://c19hcc.org/resources/npi-dashboard/ (2020).
71. Wu, X., Nethery, R. C., Sabath, B. M., Braun, D. & Dominici, F. Exposure to air pollution and COVID-19 mortality in the united states: a nationwide cross-sectional study. Preprint at https://www.medrxiv.org/content/10.1101/2020.04.05.20054502v2 (2020).
72. Bigquery public datasets. https://cloud.google.com/bigquery/public-data (2020).
73. Biddison, E. et al. Scarce resource allocation during disasters: a mixed-method community engagement study. *Chest* **153**, 187–195 (2018).
74. COVID-19 search trends symptoms dataset. https://pantheon.corp.google.com/marketplace/product/bigquery-public-datasets/covid19-search-trends?project=covid-forecasting-272503&folder=&organizationId= (2021).
75. Nunan, D. et al. Covid-19 symptoms tracke. https://www.cebm.net/covid-19/covid-19-signs-and-symptoms-tracker/ (2020).
76. Covid-19 community mobility reports. https://www.google.com/covid19/mobility/ (2020).
77. Aktay, A. et al. Google COVID-19 community mobility reports: anonymization process description (version 1.1). https://europepmc.org/article/PPR/PPR272259 (2020).
78. Policies of the office of the prime minister of japan and his cabinet. https://japan.kantei.go.jp/ongoingtopics/index.html (2021).
79. Gillam, M. Japan's response to the coronavirus pandemic. https://www.jlgc.org/04-28-2020/8414/ (2020).
80. Barkay, N. et al. Weights and methodology brief for the COVID-19 symptom survey by University of Maryland and Carnegie Mellon University, in partnership with Facebook. Preprint at https://arxiv.org/abs/2009.14675 (2020).
81. The Statistics Bureau of Japan. *Natl Census* http://www.stat.go.jp/data/kokusei/2015/ (2015).
82. The Statistics Bureau of Japan. The 66th Japan statistical yearbook 2017. http://www.stat.go.jp/data/nenkan/66nenkan/02.html (2015).
83. Organisation for Economic Cooperation and Development. *OECD.stat* https://stats.oecd.org/ (2000).
84. Ministry of Health, Labor, and Welfare. Handbook of health and welfare statistics. https://www.mhlw.go.jp/english/database/db-hh/2-2.html (2019).
85. Ministry of Health, Labor, and Welfare. Survey on medical treatment status and number of beds accepted by inpatients. https://www.mhlw.go.jp/stf/seisakunitsuite/newpage_00023.html (2020).
86. Ministry of Health, Labor, and Welfare. 2012 national health and nutrition survey report. https://www.mhlw.go.jp/bunya/kenkou/eiyou/h24-houkoku.html (2012).
87. National Tax Agency. Statistics on imposition of liquor tax 2016. https://www.nta.go.jp/taxes/sake/tokei/mokuji.htm (2016).
88. Infectious Disease Surveillence Center. Statistics on imposition of liquor tax 2016. http://idsc.nih.go.jp/idwr/sokuho/index.html (2010).
89. Ministry of Health, Labor, and Welfare. Comprehensive survey of living conditions. https://www.mhlw.go.jp/toukei/list/20-21.html (2007).
90. Kirkcaldy, R. D., King, B. A. & Brooks, J. T. COVID-19 and postinfection immunity: limited evidence, many remaining questions. *J. Am. Med. Assoc.* **323**, 2245–2246 (2020).
91. Meng, X. & Chen, L. The dynamics of a new sir epidemic model concerning pulse vaccination strategy. *Appl. Mathematics Comput.* **197**, 582 – 597 (2008).
92. Different COVID-19 vaccines. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html (2021).
93. Hens, N. et al. Seventy-five years of estimating the force of infection from current status data. *Epidemiol. Infect.* **1386**, 802–12 (2010).
94. van den Driessche, P. & Watmough, J. Further notes on the basic reproduction number. in *Lecture Notes in Mathematics, LNM* Chapter 6, Vol. 1945, 159–178, Eds. F. Brauer, P. van den Driessche & J. Wu (Springer, 2008).
95. Fda briefing document moderna covid-19 vaccine. https://www.fda.gov/media/144434/download/ (2021).
96. Hastie, T. & Tibshirani, R. Generalized additive models. *Statist. Sci.* **1**, 297–310 (1986).
97. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (Association for Computing Machinery, 2016).
98. Bengio, S., Vinyals, O., Jaitly, N. & Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* - Volume 1 (NIPS'15, 2015).
99. Golovin, D. et al. Google vizier: a service for black-box optimization. in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1487–1495 (ACM, 2017).
100. Wen, R., Torkkola, K., Narayanaswamy, B. & Madeka, D. A multi-horizon quantile recurrent forecaster. *Proceedings of the 31st Conference on Neural Information Processing Systems*, (NeurIPS, 2017).
101. Mariano, R. S. & Diebold, F. X. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **13**, 253 (1995).
102. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
103. Kwiatkowski, D., Phillips, P. C., Schmidt, P. & Shin, Y. 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? *J. Econometrics* **54**, 159–178 (1992).
104. Labgold, K. et al. Widening the gap: greater racial and ethnic disparities in COVID-19 burden after accounting for missing race/ethnicity data. Preprint at https://www.medrxiv.org/content/10.1101/2020.09.30.20203315v1.full (2020).
105. Baqui, P., Bica, I., Marra, V., Ercole, A. & van Der Schaar, M. Ethnic and regional variations in hospital mortality from COVID-19 in brazil: a cross-sectional observational study. *Lancet Global Health* **8**, e1018–e1026 (2020).
106. Aldridge, R. W. et al. Black, Asian and minority ethnic groups in England are at increased risk of death from COVID-19: indirect standardisation of NHS mortality data. *Wellcome Open Res.* **5**, 88 (2020).
107. Tai, D. B. G., Shah, A., Doubeni, C. A., Sia, I. G. & Wieland, M. L. The disproportionate impact of COVID-19 on racial and ethnic minorities in the united states. *Clin. Infect. Dis.* **16**, 703–706 (2020).
108. Laurencin, C. T. & McClinton, A. The COVID-19 pandemic: a call to action to identify and address racial and ethnic disparities. *J. Racial Ethnic Health Disparities* **7**, 398–402 (2020).
109. Murata, C. et al. Barriers to health care among the elderly in Japan. *Int. J. of Env. Res. Pub. Health* **7**, 1330–1341 (2010).
110. Sen, P. K. Estimates of the regression coefficient based on Kendall's tau. *J. Am. Stat. Assoc.* **63**, 1379–1389 (1968).
111. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS'17, 2017).
112. Kendall, A & Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS'17, 2017).

113. Malinin, A. & Gales, M. Predictive uncertainty estimation via prior networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (NIPS'18, 2018).
114. Ovadia, Y. et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, (2019).
115. Dusenberry, M. W. et al. Analyzing the role of model uncertainty for electronic health records. In *Proc. of the ACM Conference on Health, Inference, and Learning (CHIL '20)* 204–213 (Association for Computing Machinery, NewYork, NY, USA, 2020). https://doi.org/10.1145/3368555.3384457.
116. Filos, A. et al. Benchmarking Bayesian deep learning with diabetic retinopathy diagnosis. Preprint at https://arxiv.org/abs/1912.10481 (2019).
117. van Amersfoort, J., Smith, L., Teh, Y. W. & Gal, Y. Uncertainty estimation using a single deep deterministic neural network. *Int. Conf. Mach. Learn*. (2020).
118. Brauer, F., Castillo-Chavez, C. & Feng, Z. *Mathematical Models in Epidemiology*, Vol. 69 (Springer, 2019).

ACKNOWLEDGEMENTS

This work and the collection of data were approved by Google. Only public data were used, including aggregated, anonymized mobility data/reports (https://www.google.com/covid19/mobility/). As part of this research, we performed a thorough ethics review to inform our methods and reporting. No extramural funding was used for this project. We thank all those in the USA and Japan who access and view our public forecasts. We would also like to thank Afraz Mohammad, Alexander Titus, Andrew Max, Andrew Moore, Anne Mottram, Brian Kang, Burak Gokturk, Chris Haas, David Feinberg, Dave Steiner, David Parish, Derrick Miller, Donny Cheung, Evgeniy Gabrilovich, Garth Graham, George Teoderci, Greg Wellenius, Hanchao Liu, Heiga Zen, Helen Wang, Hootan Nakhost, Jacqueline Shreibati, Jacqueline Shreibati, Jasmin Repenning, Jeremy Kubica, Junichi Kawai, Kaho Kobayashi, Karen DeSalvo, Karen Ouk, Karthik Ramasamy, Kathryn Rough, Kevin Murphy, Lukman Ramsey, Mario Callegaro, Martin Nikoltchev, Matt Hoffman, Matthew Siegler, Meg Mitchell, Michael Howell, Nanako Yamaguchi, Natalie Wei, Nenad Tomasev, Ola Rozenfeld, Oriol Vinyals, Oscar Wahltinez, Peter Battaglia, Peter Fitzgerald, Priya Rangaswamy, Rand Xie, Rif A. Saurous, Ron Bodkin, Ryu Hirayama, Sakura Tominaga, Sandhya Patil, Shane Glass, Shinya Saito, Takahide Kato, Tin-yun Ho, Tomohiko Kikuchi, Tulsee Doshi, Vishal Karande, Yash Soothalia, Yong Li, and Yun Liu. Finally, we thank the many administrators, engineers, researchers, and physicians who have worked on the data collection, and the rest of the Google Cloud and Google Japan teams for their support, ideas, and encouragement.

AUTHOR CONTRIBUTIONS

S.A. and T.P. initiated the project. S.A. and J.S. initiated the project for Japan. S.A., J.S., R.S., J.Y., J.R.L., L.T., M.W.D., N.Y., K.P. and H.K. created the dataset. S.A., J.S., R.S., J.Y., J.R.L., L.T., M.W.D., N.Y., K.P., A.E., J.E., E.K., I.J., C.L., B.L., J.M., V.M., S.S., M.S., A.S., L.Z., K.C. and T.P. contributed to software engineering. S.A., J.S., R.S., J.Y., J.R.L., L.T., M.W.D., N.Y., K.P., D.S., H.K. and T.P. analyzed the results. S.A., J.S., R.S., J.Y., J.R.L., L.T., M.W.D., N.Y. and T.P. contributed to the overall experimental design. S.A., R.S., J.Y., L.T., M.W.D., N.Y. and T.P. designed the model architecture. J.R.L., T.T., D.Y., S.N. and H.M. contributed expertise in epidemiology and clinical medicine. J.S. and M.W.D. contributed to experiments into model uncertainty. S.A., J.S., R.S., J.Y., L.T., M.W.D., N.Y. and T.P. contributed to analyzing model performance. S.A., J.S., J.R.L., L.T., M.W.D., N.Y. and T.P. contributed to fairness analyses. S.A. and R.S. contributed to ablation experiments. S.A., J.S., R.S., J.Y., J.R.L. and T.P. contributed to experiments exploring counterfactual scenarios. S.A. and R.S. contributed to analyzing the importance of individual features. S.A., J.S., J.R.L., H.K. and T.P. managed the project. S.A., J.S., R.S., M.W.D., N.Y. and J.R.L. wrote the paper.

COMPETING INTERESTS

H.M. and S.N. are recipients of a Google.org Fellowship grant. The remaining authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-021-00511-7.

Correspondence and requests for materials should be addressed to Sercan Ö. Arık.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.