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

Joseph Ledsam (jledsam@google.com)
Google https://orcid.org/0000-0001-9917-7196

Sercan Arik
Google Cloud AI

Joel Shor
Google Japan

Rajarishi Sinha
Google Cloud AI https://orcid.org/0000-0001-9157-674X

Jinsung Yoon
Google Cloud AI

Long Le
Google Cloud AI

Michael Dusenberry
Google Cloud AI

Nate Yoder
Google Cloud AI

Kris Popendorf
Google Japan

Arkady Epshteyn
Google Cloud AI

Johan Euphrosine
Google Japan

Elli Kanal
Google Cloud AI

Isaac Jones
Google Cloud AI

Chun-Liang Li
Google Cloud AI

Beth Luan
Google Cloud AI

Joe Mckenna

Google Cloud AI

Vikas Menon
Google Cloud AI

Shashank Singh
Google Cloud AI

Mimi Sun
Google Health

Ashwin Sura Ravi
Google Cloud AI

Leyou Zhang
Google Cloud AI

Dario Sava
Google Cloud AI

Hiroki Kayama
Google Japan

Thomas Tsai
Harvard School of Public Health

Daisuke Yoneoka
School of Medicine, Keio University; Graduate School of Public Health, St Luke's International University

Shuhei Nomura
University of Tokyo

Hiroaki Miyata
School of Medicine, Keio University; Graduate School of Medicine, The University of Tokyo

Tomas Pfister
Google Cloud AI

Biological Sciences - Article

Keywords: National Mean Absolute Percentage Error, Population Behavior, Demographic Differences, Policy Interventions, Counterfactual Simulations

Posted Date: March 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-312419/v1

Version of Record: A version of this preprint was published at npj Digital Medicine on October 8th, 2021. See the published version at https://doi.org/10.1038/s41746-021-00511-7.

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

Sercan Ö. Arık[1,*,+], Joel Shor[2,+], Rajarishi Sinha[1,+], Jinsung Yoon[1,+], Joseph R. Ledsam[2,+], Long T. Le[1], Michael W. Dusenberry[1], Nate 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], Hiroki Kayama[2], Thomas Tsai[4], Daisuke Yoneoka[5,6], Shuhei Nomura[5,7], Hiroaki Miyata[5,8], and Tomas Pfister[1]

[1]Google Cloud AI, USA

[2]Google, Japan

[3]Google Health, USA

[4]Harvard School of Public Health, USA

[5]Department of Health Policy and Management, School of Medicine, Keio University, Tokyo, Japan

[6]Division of Biostatistics and Bioinformatics, Graduate School of Public Health, St Luke's International University, Tokyo, Japan

[7]Department of Global Health Policy, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[8]Department of Healthcare Quality Assessment, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[*]soarik@google.com

[+]these authors contributed equally to this work

ABSTRACT

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 and we present an international, prospective evaluation of our models' performance across all states and counties in the USA and prefectures in Japan. National mean absolute percentage error (MAPE) for predicting COVID-19 associated deaths before and after prospective deployment remained consistently <2% (US) and <10% (Japan). Average statewide (US) and prefecture wide (Japan) MAPE was 6% and 26% respectively (14% when looking at prefectures with more than 50 deaths). We show that our models perform well even during periods of considerable change in population behavior, and that it is 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.

Introduction

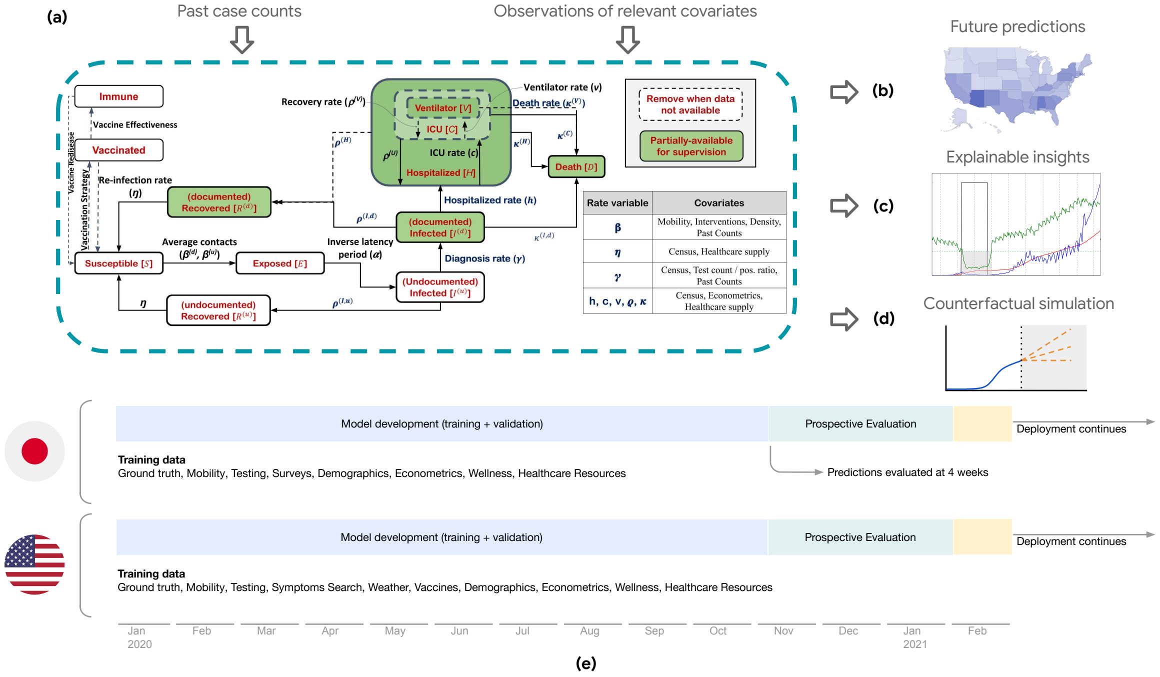
Predicting the spread of infectious diseases is an essential component of public health management. Forecasts have contributed to resource allocation and control measures in past epi- and pandemics such as influenza[1] and Ebola[2]. Most recently such models have shown promise during the COVID-19 pandemic[3,4] 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[3,4,15–18]. 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 need 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[19,20].

To address these challenges, we introduced an accurate, generalizable, 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[21], 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.



(a)
Past case counts
Observations of relevant covariates
Immune
Vaccine Effectiveness
Vaccinated
Vaccine Redisease
Vaccination Strategy
Re-infection rate
(η)
(documented) Recovered [R^(d)]
Susceptible [S]
Average contacts
(β^(d), β^(u))
Exposed [E]
Inverse latency period (α)
(Undocumented) Infected [I^(u)]
η
(undocumented) Recovered [R^(u)]
ρ^(I,u)
(documented) Infected [I^(d)]
ρ^(I,d)
ρ^(H)
Diagnosis rate (γ)
Hospitalized rate (h)
Recovery rate (ρ^(V))
Ventilator [V]
ICU [C]
ρ^(U)
ICU rate (c)
Hospitalized [H]
Ventilator rate (v)
Death rate (κ^(V))
κ^(C)
κ^(H)
κ^(I,d)
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
Future predictions
(b)
Explainable insights
(c)
Counterfactual simulation
(d)
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)

Figure 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[22, 23]. 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 Japan at the prefecture level. (c) Interpretable elements, including predictions for the effective R number and force of infection provide explainable and actionable insights. (d) Simulations of counterfactual scenarios can be used to estimate the 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.

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[22] (Figure 1). Our models are optimised for prediction of COVID-19 associated deaths, and are trained on static and time-series data (see Methods).

A prospective evaluation of forecasting accuracy

To evaluate our framework, we conducted a prospective observational study over eight weeks in the USA and Japan. Predictions were made daily, each looking 4 weeks into the future (Figure 2). Our primary analysis is based on the absolute percentage error (APE - see Methods) in the predicted number of COVID-19 associated deaths, and our secondary analyses included confirmed cases. For the USA, the availability of appropriate data also allowed the prediction of hospitalizations, intensive care (ICU) admissions and admissions requiring mechanical ventilation.

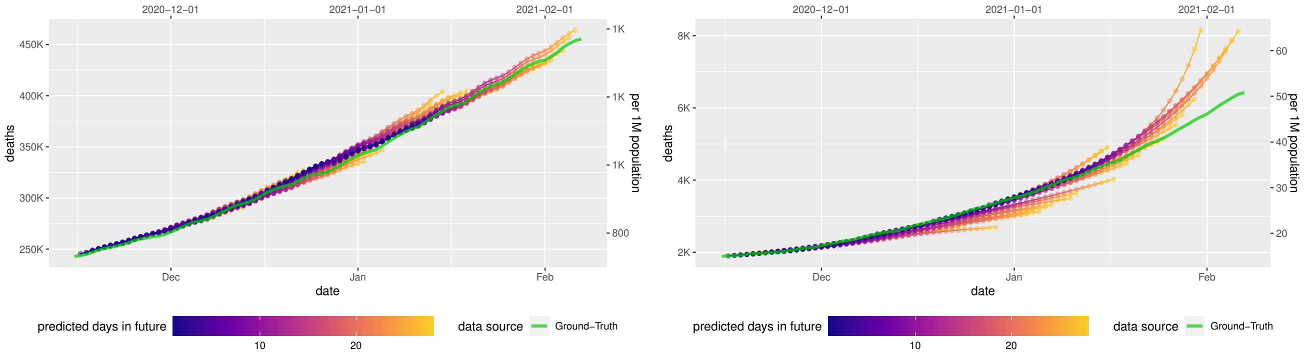
During the prospective period, across the USA as a whole, the framework achieved an aggregate absolute percentage error (AAPE) of 1.4% (95% CI [1.1%, 1.6%]) for deaths. The framework predicted confirmed cases, hospitalizations, intensive care unit (ICU) admissions, and admissions requiring mechanical ventilation with AAPEs of 9.20% (95% CI [8.3%, 10.2%]), 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 state- and county-level predictions. When evaluating at state level and averaging across all locations, the framework achieves mean absolute percentage error (MAPE) for deaths and confirmed cases of 5.4% [5.1%, 5.6%] and 9.2% [8.2%, 10.1%], respectively. At county level, MAPE for deaths and confirmed cases were 25.1% [23.1%, 27.0%] and 12.8% [11.5%, 14.1%], respectively. Predictions of deaths achieved a APE <10% or AE <100 for 43/51 states and 2585/3006 counties, and for confirmed cases 34/51 states and 1647/3006 counties (Supplementary Tables 1 & 2).

We can adjust the overall accuracy of our forecasts to fit different use-cases by only releasing the most confident predictions. Our framework is well calibrated: uncertainty correlates with 28-day forecast accuracy (Supplementary Section 7). Thus we can withhold less confident predictions, observing a 25% reduction in MAPE for Japan deaths by only releasing the most confident 50% of predictions (Extended Data Figure 1).

We compare our framework with alternatives. We evaluate significance with a two-sided Diebold-Mariano (DM) test using MAE (Supplementary Table 13). For deaths, DM statistics are negative, indicating our framework has a lower MAE than all alternative models. This difference is statistically significant in 12/30 comparisons. Using MAPE, ’COVIDhub-ensemble’ - a combined forecast that includes our frameworks predictions - has a negative DM statistic and a slightly lower MAPE, but this difference is not statistically significant (Supplementary Table 14). Similarly for cases prediction, for which our model was not optimized, there was no significant difference for MAE. Using MAPE, two models were significantly better. In addition to MAPE, we also compare models using the weighted interval score [24,25] and find our framework consistently ranks top or in the top-5 models (Extended Data Figures 3 4, Supplementary Section 2).

For Japan, we report AAPEs for deaths and confirmed cases, 4 weeks ahead of time, 9.8% (95% CI [7.4%, 12.2%]) and 9.1% (95% CI [5.7%, 12.5%] respectively. Data were not available on hospitalizations, ICU admissions and admissions requiring mechanical ventilation. At prefecture-level, MAPEs for deaths and confirmed cases were 25.9% [24.4%, 27.5%] and 21.4% [19.5%, 23.4%] respectively. The number of prefectures with an APE <10% or AE <10 for deaths and confirmed cases were 38/47 and 14/47 respectively (Supplementary Tables 3 & 4).

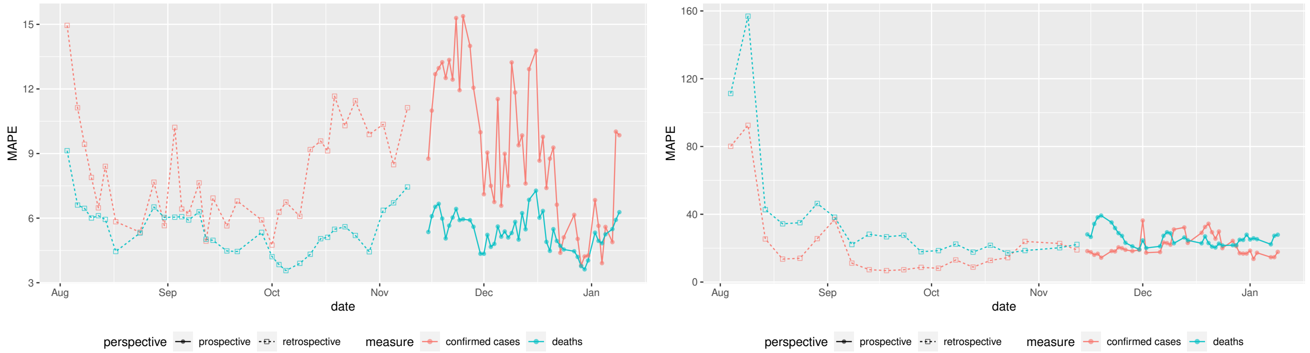
In addition to evaluating our framework prospectively, we also show retrospective evaluations for dates before the prospective study began. Retrospective performance was achieved by training the model using data up to a particular prediction date, and evaluating 4 weeks after the prediction date. Though there is no leakage of data from the future dates, the framework uses the most recent version that includes any corrections made to previous data. Our comparison shows that the MAPE during the



(a)

(b)

Figure 2. Prospective forecasts for the USA 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.



(a)

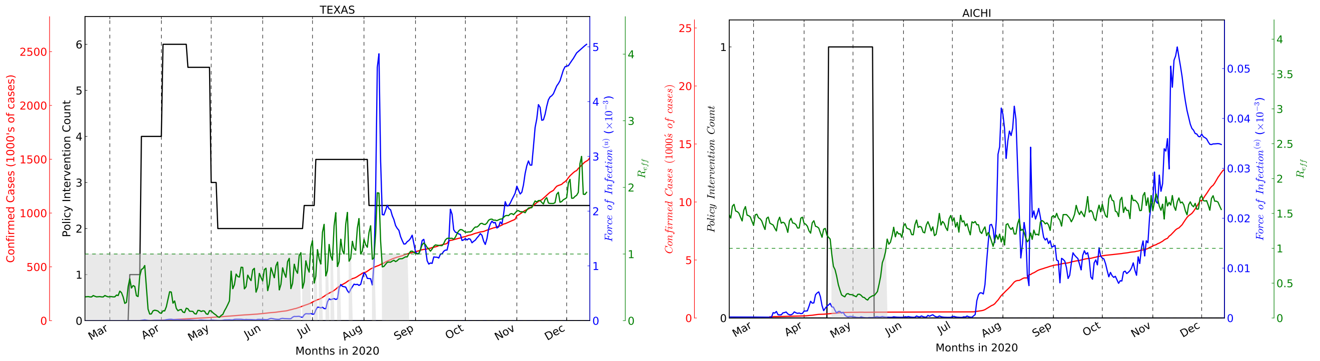
(b)

Figure 3. Retrospective and prospective 28-day MAPE over time. Performance over time is shown for the (a) state-level USA models (b) prefecture-level Japan model. 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).

prospective period was at most 1.3% above the MAPE for the retrospective period for both deaths and confirmed cases in both the USA and Japan (Figure 3).

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. However, different settings may benefit from other prediction horizons (see Extended Data Figure 2).

COVID-19 disproportionately affects certain demographic subpopulations[26–30]. 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, after accounting for confounding variables, only small correlations remained for most subgroups suggesting that the errors are not associated with the demographic variables of race, gender, population density or income (Extended Data Table 1, Supplementary Section 6).



(a)

(b)

Figure 4. Interpretable model outputs. 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 numbers of COVID-19 associated deaths respectively. Confirmed case counts and number of Non-Pharmaceutical Interventions (NPIs) are plotted on the left Y-axis, and $F^{(u)}$ (see Eq. 1) and the R_{eff} (see Eq. 3) are plotted on the right Y-axis. For $R_{eff} < 1$ (shaded grey regions below the horizontal dotted line), dynamics are tending towards the Disease-Free Equilibrium (DFE)[31]. These areas often overlap with the dates when multiple NPIs are imposed.

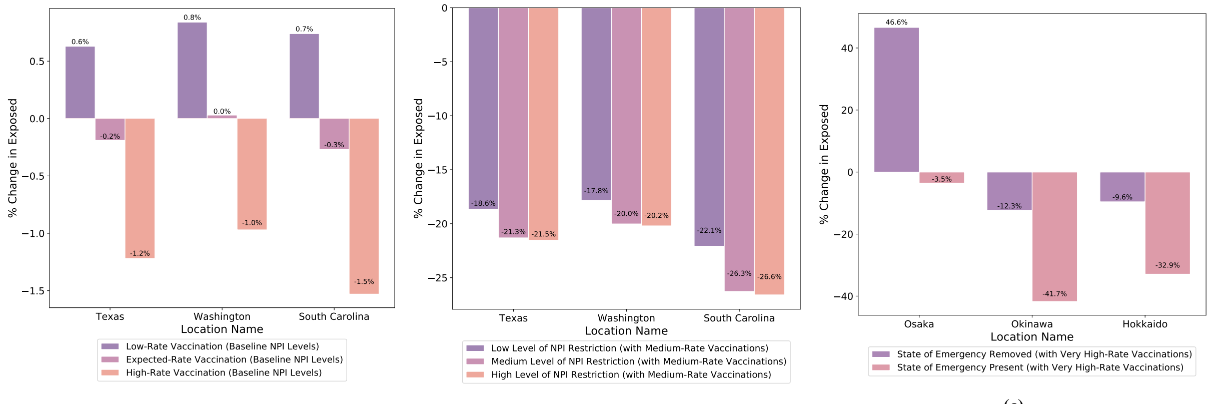
Using the framework to understand the COVID-19 pandemic

Modeling compartments and the transitions between them explicitly allows predictions of how connected compartments change over time. This offers insights into disease dynamics, including estimates for the effective R number (R_{eff}), and the force of infection ($F^{(u)}$, the rate at which susceptible individuals acquire the disease). Figure 4 demonstrates this for Texas, USA and Aichi, Japan respectively (for all other locations see Supplementary Figures 18-28. We observe that non-pharmaceutical interventions (NPIs, such as mask mandates and mobility restrictions[32]) 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 $F^{(u)}$.

Additionally the effect of individual features on the transition rates (modeled by encoders, see Methods) provides insights on the relative contributions of each feature (Extended Data Table 2, Supplementary Section 4). Internet search trends, survey results for COVID-like symptoms, and weather trends were most strongly associated with fitted contact rate. The encoder weights may also be helpful in comparing NPIs: of the seven considered for the USA, closing schools ranks higher than others, suggesting its relative contribution to reducing COVID-19 spread may be greater.

Simulating the effects of interventions

Our framework can be used to predict the effect of interventions including NPIs and vaccinations. Overriding NPI features provides 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 cases and deaths (Supplementary Table 22) as well as evaluating on simulated data



(a) **(b)** **(c)**

Figure 5. 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 baseline exposed individual counts are 69694, 67591 and 63742 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, more than 17 % across the three states. 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 very high rate vaccination (2% population/day), and considering the cases of applying or removing the State of Emergency. Here, the baseline exposed individual counts are 5779, 3838 and 3253 for Osaka, Okinawa and Hokkaido respectively. Applying the the State of Emergency is observed to be highly 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.

(Supplementary Section 3).

With counterfactual analysis, consistent with the findings from feature importance, we found that school closures are associated with the highest reduction in predicted exposed counts among all NPIs. Joint application of multiple NPIs are observed to be much more effective than each individual (Extended Data Tables 3, 4 and 5). We also find that a 7-day delay in applying all NPIs reduced predicted cases by 45% for Japan (Extended Data Table 8). Maintaining NPIs during vaccination drives results 9% and 9.3% reductions in predicted cases and deaths. When increasing the vaccination rate to 1% of the population per day is considered, 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 visible over longer time horizons, beyond 4 weeks and 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 high vaccination rate seems highly beneficial (Figure 5(c), Extended Data Table 7, Supplementary Figure 17)).

Use cases and the impact of our framework

Our forecasts are released publicly, and thus are available to a wide range of organisations to whom the information may aid decision making[33]. 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 44.

Discussion

We present and evaluate prospectively an AI-augmented approach to epidemiology that forecasts at a state, county and prefecture level, and provides 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 decision making.

The forecasts are publicly available[33], and have been adopted alongside other information by a number of public and private organisations, 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 be used to suggest locations where undocumented infections are rapidly increasing, and where increasing testing may be beneficial.

Our framework can help understand the potential consequences of public health decisions around NPIs and vaccination with counterfactual analysis. Via modifications to the proposed compartmental model, it is possible to model the efficacy for different vaccine regimens as new vaccines and strategies become available, ensuring the framework remains 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 survey conducted on organisations using the framework included academic and government organisations that had actively used this counterfactual analysis capability in their decision making.

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. Firstly, cases in Japan are skewed towards a small number of prefectures. This means the model training is dominated by a small number of locations. The uneven distribution also means super-spreaders may be more dominant in population-dense areas, which decreases predictability for Japan compared with the US. Secondly, there was less data to learn from due to fewer COVID-19 cases, and Japan ICU and hospitalization data were unavailable for modeling. More generally, data quality was not always consistent, including errors such as reporting delays and incorrect data. 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]. Finally, our models were optimized for predicting COVID-19 associated deaths, our primary analysis. It is possible that performance for case prediction could be improved if the models were instead optimised 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 on the remaining predictions. As each prediction provides an estimate for 4 weeks ahead, the impact 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[34], 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 available[35]. 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[36–40], compartmental models[22,23,41–45], and machine learning[46–51]. Recent work has modeled the impact of NPIs such as travel restrictions[52] in the US[53] and Europe[54]. However, these studies have been limited to integration of one or two features, often with judiciously-designed functional forms. Standard compartmental models fit the COVID-19 pandemic data poorly. By modeling static and time-varying features in conjunction, learning their associations from data in an end-to-end way, our model improves performance while bringing explainable insights[21]. Conversely, several recent publications have attempted direct modeling from features[55–57]. In the absence of high-quality and large-scale historical data, these methods under-perform as they lack an inductive bias coming from epidemiological basis, which also limits interpretability and thus applicability. 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. Compared to others, our framework is consistently accurate, gives forecasts for more compartments, offers explainability, reliable counterfactual analysis, and is generalizable to both higher granularity and other countries.

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. USA 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 on past events and simulated data constitute only partial solutions (Supplementary Section 3) as prospective evaluations of counterfactual outcomes present feasibility challenges. 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 is 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. For these reasons, 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 remains a global emergency. As governments, business and individuals face substantial new challenges ahead 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 counts not only possible, they are an essential and growing part of public health.

References

1. Shaman, J. & Karspeck, A. Forecasting seasonal outbreaks of influenza. *Proc. Natl. Acad. Sci.* **109**, 20425–20430, DOI: 10.1073/pnas.1208772109 (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. *medRxiv* DOI: 10.1101/2020.08.19.20177493 (2020).

4. Reiner, R. C. *et al.* Modeling COVID-19 scenarios for the united states. *Nat. Medicine* (2020).

5. Bedford, J. *et al.* COVID-19: towards controlling of a pandemic. *The 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. *The Lancet Glob. Heal.* **8**, e480 (2020).

8. Bonaccorsi, G. *et al.* Economic and social consequences of human mobility restrictions under COVID-19. *Proc. Natl. Acad. Sci.* **117**, 15530–15535 (2020).

9. Borio, C. The COVID-19 economic crisis: dangerously unique. *Bus. Econ.* **55**, 181–190, DOI: 10.1057/s11369-020-00184-2 (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 Heal.* **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. *PsyArXiv:10.31234/osf.io/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 (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, DOI: 10.1145/3233547.3233667 (Association for Computing Machinery, New York, NY, USA, 2018).

15. Fu, X. Global analysis of daily new COVID-19 cases reveals many static-phase countries including us and uk potentially with unstoppable epidemics. *medRxiv* (2020).

16. Long, Y.-S. *et al.* Quantitative assessment of the role of undocumented infection in the 2019 novel coronavirus (COVID-19) pandemic. *arXiv:2003.12028* (2020).

17. Chang, S. *et al.* Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* (2020).

18. Rodriguez, A. *et al.* Deepcovid: An operational deep learning-driven framework for explainable real-time COVID-19 forecasting. *medRxiv* (2020).

19. 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
Accessed: 2021-01-31.

20. Japanese association of infectious diseases : Recommendations for COVID-19 vaccine. https://www.kansensho.or.jp/modules/guidelines/index.php?content_id=43. Accessed: 2021-01-31.

21. Arik, S. O. *et al.* Interpretable sequence learning for COVID-19 forecasting (2020). 2008.00646.

22. Blackwood, J. C. & Childs, L. M. An introduction to compartmental modeling for the budding infectious disease modeler. *Lett. Biomath.* **5**, 195–221 (2018).

23. Smith, D. & Moore, L. The sir model for spread of disease (2004).

24. Bracher, J., Ray, E. L., Gneiting, T. & Reich, N. G. Evaluating epidemic forecasts in an interval format (2020). 2005.12881.

25. Gneiting, T. & Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **102**, 359–378, DOI: 10.1198/016214506000001437 (2007).

26. Yancy, C. W. COVID-19 and African Americans. *JAMA* **323**, 1891–1892 (2020).

27. Webb Hooper, M., Nápoles, A. M. & Pérez-Stable, E. J. COVID-19 and Racial/Ethnic Disparities. *JAMA* **323**, 2466–2467 (2020).

28. Chowkwanyun, M. & Reed, A. L. Racial health disparities and COVID-19 — caution and context. *New Engl. J. Medicine* **383**, 201–203 (2020).

29. 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. *The Lancet* **395**, 1673–1676 (2020).

30. Henning-Smith, C., Tuttle, M. & Kozhimannil, K. B. Unequal distribution of COVID-19 risk among rural residents by race and ethnicity. *The J. rural health* 10.1111/jrh.12463 (2020).

31. Brauer, F., Castillo-Chavez, C. & Feng, Z. *Mathematical Models in Epidemiology*, vol. 69 (Springer, New York, NY, 2019).

32. Non-pharmaceutical interventions. Accessed 2020-02-17.

33. COVID-19 public forecasts. https://pantheon.corp.google.com/marketplace/product/bigquery-public-datasets/covid19-public-forecasts. Accessed: 2021-02-18.

34. Demirgüç-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 Res. Work. Pap.* (2020).

35. Wahltinez, O. *et al.* COVID-19 open-data: curating a fine-grained, global-scale data repository for sars-cov-2 (2020). Work in progress.

36. Capasso, V. Reaction-diffusion models for the spread of a class of infectious diseases. In Neunzert, H. (ed.) *Proceedings of the Second European Symposium on Mathematics in Industry*, vol. 3, 181–194 (Springer, Dordrecht, 1988).

37. Hunter, E., Namee, B. M. & Kelleher, J. An open-data-driven agent-based model to simulate infectious disease outbreaks. *PLoS ONE* **13** (2018).

38. 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. *medRxiv* (2020).

39. GrowthRate model from Los Alamos National Laboratory. https://COVID-19.bsvgateway.org/#link%20to%20forecasting%20site. Accessed: 2020-06-04.

40. White, S. H., del Rey, A. M. & Sánchez, G. R. Modeling epidemics using cellular automata. *Appl. Math. Comput.* **186**, 193–202 (2006).

41. Kermack, W. O. & McKendrick, A. G. Contributions to the mathematical theory of epidemics—i. *Proc. Royal Soc.* **115A**, 700–721 (1927).

42. Grand Rounds. COVID-19 forecasting: Fit to a curve or model the disease in real-time? (2020). https://grandrounds.com/blog/COVID-19-forecasting-fit-to-a-curve-or-model-the-disease-in-real-time/, Last accessed on 2020-05-29.

43. Li, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science* **368**, 489–493 (2020).

44. Koller, D. & Friedman, N. *Probabilistic graphical models: principles and techniques* (MIT press, 2009).

45. Osthus, D., Hickmann, K. S., Caragea, P. C., Higdon, D. & Del Valle, S. Y. Forecasting seasonal influenza with a state-space SIR model. *The annals applied statistics* **11**, 202 (2017).

46. Greydanus, S., Dzamba, M. & Yosinski, J. Hamiltonian neural networks. *arXiv:1906.01563* (2019).

47. Cranmer, M. *et al.* Lagrangian neural networks (2020). 2003.04630.

48. Lutter, M., Ritter, C. & Peters, J. Deep lagrangian networks: Using physics as model prior for deep learning. *arXiv:1907.04490* (2019).

49. Iclr 2020 workshop on integration of deep neural models and differential equations. http://iclr2020deepdiffeq.rice.edu/. Accessed: 2020-06-04.

50. Lim, B., Arik, S. O., Loeff, N. & Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *arXiv:1912.09363* (2019).

51. Williams, R. J. & Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1**, 270–280 (1989).

52. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400, DOI: 10.1126/science.aba9757 (2020).

53. Pei, S., Kandula, S. & Shaman, J. Differential effects of intervention timing on COVID-19 spread in the united states. *medRxiv* (2020).

54. 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).

55. Venna, S. R. *et al.* A novel data-driven model for real-time influenza forecasting. *IEEE Access* **7**, 7691–7701 (2019).

56. Yang, Z. *et al.* Modified seir and ai prediction of the epidemics trend of COVID-19 in china under public health interventions. *J. Thorac. Dis.* **12**, 165–174 (2020).

57. Wang, L., Chen, J. & Marathe, M. Tdefsi: Theory guided deep learning based epidemic forecasting with synthetic information (2020). 2002.04663.

58. Ensheng Dong, H., Du & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infect. Dis.* **20**, 533–534 (2020).

59. Covid-Tracking. The covid tracking project (2020).

60. Bryant, P. & Elofsson, A. Estimating the impact of mobility patterns on COVID-19 infection rates in 11 european countries. *medRxiv* DOI: 10.1101/2020.04.13.20063644 (2020).

61. Warren, M. S. & Skillman, S. W. Mobility changes in response to COVID-19. *arXiv:2003.14228 [cs.SI]* (2020).

62. Realtime tracking of state-wide npi implementations. https://c19hcc.org/resources/npi-dashboard/. Accessed: 2020-06-04.

63. 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. *medRxiv* (2020).

64. Bigquery public datasets. https://cloud.google.com/bigquery/public-data. Accessed: 2020-06-04.

65. Biddison, E. *et al.* Scarce resource allocation during disasters: A mixed-method community engagement study. *Chest* **153**, 187–195 (2018).

66. 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=. Accessed: 2021-02-01.

67. Nunan, D. *et al.* Covid-19 symptoms tracker (2020).

68. Covid-19 community mobility reports (2020).

69. Aktay, A. *et al.* Google COVID-19 community mobility reports: Anonymization process description (version 1.1) (2020).

70. Policies of the office of the prime minister of japan and his cabinet. https://japan.kantei.go.jp/ongoingtopics/index.html. Accessed: 2021-02-01.

71. Gillam, M. Japan's response to the coronavirus pandemic. https://www.jlgc.org/04-28-2020/8414/ (2020). Accessed: 2021-02-08.

72. 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 (2020).

73. The Statistics Bureau of Japan. National census (2015).

74. The Statistics Bureau of Japan. The 66th Japan statistical yearbook 2017 (2015).

75. Organisation for Economic Cooperation and Development. OECD.stat (2000).

76. Ministry of Health, Labor, and Welfare. Handbook of health and welfare statistics (2019).

77. Ministry of Health, Labor, and Welfare. Survey on medical treatment status and number of beds accepted by inpatients (2020).

78. Ministry of Health, Labor, and Welfare. 2012 national health and nutrition survey report (2012).

79. National Tax Agency. Statistics on imposition of liquor tax 2016 (2016).

80. Infectious Disease Surveillence Center. Statistics on imposition of liquor tax 2016 (2010).

81. Ministry of Health, Labor, and Welfare. Comprehensive survey of living conditions (2007).

82. Kirkcaldy, R. D., King, B. A. & Brooks, J. T. COVID-19 and postinfection immunity: Limited evidence, many remaining questions. *JAMA* (2020).

83. Meng, X. & Chen, L. The dynamics of a new sir epidemic model concerning pulse vaccination strategy. *Appl. Math. Comput.* **197**, 582 – 597 (2008).

84. Different COVID-19 vaccines. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html. Accessed: 2021-01-04.

85. Hens, N. *et al.* Seventy-five years of estimating the force of infection from current status data. *Epidemiol. infection* **138 6**, 802–12 (2010).

86. van den Driessche, P. & Watmough, J. *Further Notes on the Basic Reproduction Number*, chap. 6, 159–178. Lecture Notes in Mathematics, LNM vol 1945 (Springer, Berlin, Heidelberg, 2008).

87. Fda briefing document moderna covid-19 vaccine. https://www.fda.gov/media/144434/download/. Accessed: 2021-02-15.

88. Hastie, T. & Tibshirani, R. Generalized additive models. *Stat. Sci.* **1**, 297–310, DOI: 10.1214/ss/1177013604 (1986).

89. 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, New York, NY, USA, 2016).

90. Bengio, S., Vinyals, O., Jaitly, N. & Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. *arXiv:1506.03099* (2015).

91. 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).

92. Mariano, R. S. & Diebold, F. X. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **13**, 253 (1995).

93. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. journal statistics* 65–70 (1979).

94. 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).

95. Labgold, K. *et al.* Widening the gap: greater racial and ethnic disparities in COVID-19 burden after accounting for missing race/ethnicity data. *Medrxiv* (2020).

96. 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. *The Lancet Glob. Heal.* **8**, e1018–e1026 (2020).

97. 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 research* **5** (2020).

98. 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.* (2020).

99. 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).

100. Murata, C. *et al.* Barriers to health care among the elderly in japan. *Int. journal environmental research public health* **7**, 1330–1341 (2010).

101. Sen, P. K. Estimates of the regression coefficient based on kendall's tau. *J. Am. statistical association* **63**, 1379–1389 (1968).

102. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles (2017). 1612.01474.

103. Kendall, A. & Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems* (2017). 1703.04977.

104. Malinin, A. & Gales, M. Predictive Uncertainty Estimation via Prior Networks. In *Advances in Neural Information Processing Systems* (2018). 1802.10501.

105. Ovadia, Y. *et al.* Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *arXiv:1906.02530 [cs, stat]* (2019). 1906.02530.

106. 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 (ACM CHIL)*, 204–213, DOI: 10.1145/3368555.3384457 (ACM, Toronto Ontario Canada, 2020).

107. Filos, A. *et al.* Benchmarking Bayesian Deep Learning with Diabetic Retinopathy Diagnosis. *arXiv:1912.10481 [stat.ML]* (2019). 1912.10481.

108. van Amersfoort, J., Smith, L., Teh, Y. W. & Gal, Y. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. *arXiv:2003.02037 [cs, stat]* (2020). 2003.02037.

109. 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, Santa Monica, CA, 2020).

Acknowledgements:

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., J.S., & T.P. initiated the project. S.A., J.S., R.S., J.Y., J.R.L., L.T., M.W.D., N.Y., K.P., & 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. & 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., & T.P. analysed the results. S.A., J.S., R.S., J.Y., J.R.L., L.T., M.W.D., N.Y., & T.P. contributed to the overall experimental design. S.A., R.S., J.Y., L.T., M.W.D., N.Y. & T.P. designed the model architecture. J.R.L., T.T., D.Y., S.N., & H.M. contributed expertise in epidemiology and clinical medicine. J.S., & M.W.D. contributed to experiments into model uncertainty. S.A., J.S., R.S., J.Y., L.T., M.W.D., N.Y. & T.P. contributed to analysing model performance. S.A., J.S., J.R.L., L.T., M.W.D., N.Y. contributed to fairness analyses. S.A. & R.S. contributed to ablation experiments. S.A., J.S., R.S., J.Y., J.R.L. & T.P. contributed to experiments exploring counterfactual scenarios. S.A. & R.S. contributed to analysing the importance of individual features. S.A., J.S., J.R.L., H.K. & T.P. managed the project. S.A., J.S., R.S. & J.R.L. wrote the paper.

Methods

Ethics and funding

This work, and the collection of data was approved by Google. Only public data was 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.

Study design

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

Data sources and preprocessing

In this section, we describe the datasets used for our proposed framework. The ground truth data for the compartments supervise the forecasting model via training objective functions. We also use 'static' (i.e. those with value 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 econometrics indicators. Time-varying variables such as population mobility, hospital resource usage and public policy decisions can also be important. While variables with predictive signal 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 signal 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 Tables 1 and 2 respectively).

USA model

Ground Truth for Compartments. For confirmed and death cases JHU[58] is used in our work, similar to other models[38]. They obtain the raw data from the state and county health departments. Because of the rapid progression of the pandemic, past data has often been restated, or the data collection protocols have been changed. We always use the latest version of the data

Table 1. Features used by the USA models.

Feature	Transition rates the feature is used for
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
Non-pharmaceutical 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

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[59].

Mobility. Human mobility within a region, for work and personal reasons, may have an effect on the average contact rates[60]. We use time-varying mobility indices provided by Descartes labs at both state- and county-level resolutions[61]. Descartes Labs aggregates the movement data of individual cellphone users within a region over a 24-hour 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.

Non-Pharmaceutical Interventions. Public policy decisions restricting certain classes of population movement or interaction can have a beneficial effect on restricting the progression of the disease[53], 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[62]. 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[1] reports the number of individuals over the age of 60 in USA counties. 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[63]. We use the BigQuery public dataset that comes from the USA Environmental Protection Agency (EPA) that documents historical air quality indices at the county level[2]. This static feature is encoded into the recovery rates (η), recovery ($\rho^{(I,d)}$, $\rho^{(I,u)}$, $\rho^{(H)}$, $\rho^{(C)}$,

[1] On BigQuery at c19hcc-info-ext-data:c19hcc_info_public.Kaiser_Health_demographics_by_Counties_States

[2] On BigQuery at bigquery-public-data:epa_historical_air_quality.pm10_daily_summary

$\rho^{(V)}$) and death rates ($\kappa^{(I,d)}$, κ^{H}, κ^{C}, κ^{V}), at both the state- and county-level of geographic resolution.

Econometrics. 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[64]–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[65]. 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[66] for USA states. From these symptoms we select seven[67] 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.

Weather. The Open Covid Dataset[35] 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[59]. 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 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.

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[35].

Mobility. We use 6 publicly available Google Mobility Reports[68,69] timeseries, corresponding to: retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. Each timeseries is an index referenced to a baseline value of 100 from before the pandemic. The number of unique visitors per day to places in each of the 6 categories is the raw measure. The raw measure is anonymized by adding Laplace noise. For each of the 6 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 timeseries. 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[70] that are applied by the Japanese Government on a per-prefecture basis. Local- and prefecture-level authorities in Japan have wide leeway in the interepretation of the NPI[71]. We manually map the NPI to a binary-valued timeseries. This is encoded into the contact rates ($\beta^{(d)}$, $\beta^{(u)}$).

Symptoms Survey. The Facebook Symptoms Survey dataset[72] 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 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[73] and age distributions[74] 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}).

Econometrics. We use prefecture-level per capita GDP from 2000[75] as an econometrics 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[76] and clinic beds[77], 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[78], alcohol consumption[79], past H1N1 illness[80], and smoking habits[81]. 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 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.

Corrections for locations

Missing data and preprocessing

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.

Table 2. Features used by the Japan models.

Feature	**Transition rates the feature is used for**
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

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 time-steps since the beginning of training, 22 January 2020.

Proposed compartmental model

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

- **Undocumented infected and recovered compartments**: Recent studies suggest that majority of the infected people are not detected and they dominate disease progression

  citeLi489, Fu2020undocumented, long2020quantitative (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[65] 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 demonstrate immunity to future infection. Due to reports of reinfection[82] 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[83], we introduce a new “Vaccinated" compartment, which has a transition from the “Susceptible". Approved COVID-19 vaccines have partial effectiveness[84]. 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 some approved vaccines are injected in two doses[84] and we consider those who get both doses in counting the number of vaccinated as the vaccine effectiveness numbers are quoted for two doses. Our framework also models the partial immunity between the first and second doses.

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

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. Fig. 1 describes transition rate variables used to relate the compartments, via the equations (we omit the index i for concision):

$$\begin{aligned}
S[t]-S[t-1] &= -(\beta^{(d)}I^{(d)}[t-1]+\beta^{(u)}I^{(u)}[t-1])\tfrac{S[t-1]}{N[t-1]}+\eta(R^{(d)}[t-1]+R^{(u)}[t-1])-Y[t-1],\\
E[t]-E[t-1] &= (\beta^{(d)}I^{(d)}[t-1]+\beta^{(u)}I^{(u)}[t-1])\tfrac{S[t-1]}{N[t-1]}-\alpha E[t-1],\\
I^{(u)}[t]-I^{(u)}[t-1] &= \alpha E[t-1]-(\rho^{(I,u)}+\gamma)I^{(u)}[t-1],\\
I^{(d)}[t]-I^{(d)}[t-1] &= \gamma I^{(u)}[t-1]-(\rho^{(I,d)}+\kappa^{(I,d)}+h)I^{(d)}[t-1],\\
R^{(u)}[t]-R^{(u)}[t-1] &= \rho^{(I,u)}I^{(u)}[t-1]-\eta R^{(u)}[t-1],\\
R^{(d)}[t]-R^{(d)}[t-1] &= \rho^{(I,d)}I^{(d)}[t-1]+\rho^{(H)}(H[t-1]-C[t-1])-\eta R^{(d)}[t-1],\\
H[t]-H[t-1] &= hI^{(d)}[t-1]-(\kappa^{(H)}+\rho^{(H)})(H[t-1]-C[t-1])-\kappa^{(C)}(C[t-1]-V[t-1])-\kappa^{(V)}V[t-1],\\
C[t]-C[t-1] &= c(H[t-1]-C[t-1])-(\kappa^{(C)}+\rho^{(C)}+v)(C[t-1]-V[t-1])-\kappa^{(V)}V[t-1],\\
V[t]-V[t-1] &= v(C[t-1]-V[t-1])-(\kappa^{(V)}+\rho^{(V)})V[t-1],\\
D[t]-D[t-1] &= \kappa^{(V)}V[t-1]+\kappa^{(C)}(C[t-1]-V[t-1])+\kappa^{(H)}(H[t-1]-C[t-1])+\kappa^{(I,d)}I^{(d)}[t-1],
\end{aligned}$$

Force of infection: The Force of Infection is defined as the measure of the rate at which susceptible individuals become infected[85] – for undocumented infected, formulated as:

$$F^{(u)}=\beta^{(u)}*I^{(u)}/N, \tag{1}$$

and documented infected, formulated as:

$$F^{(d)}=\beta^{(d)}*I^{(d)}/N. \tag{2}$$

Effective reproductive number: Using the Next-Generation Matrix method[86] on the proposed compartmental model, effective reproductive number can be derived as[21]:

$$R_e = \frac{\beta^{(d)}\gamma + \beta^{(u)}(\rho^{(I,d)} + \kappa^{(I,d)} + h)}{(\gamma + \rho^{(I,u)}) \cdot (\rho^{(I,d)} + \kappa^{(I,d)} + h)}. \tag{3}$$

Integration of vaccination: We consider two-dose vaccination strategy[87] and define the first-dose effectiveness function as:

$$\pi^{(1)}[\tau] = \min(\pi^{(1)}_{max}, \pi^{(1)}_{max} \cdot \tau / T^{(1)}_{\pi}), \tag{4}$$

and the second-dose effectiveness function as:

$$\pi^{(2)}[\tau] = \min(\pi^{(2)}_{max}, \pi^{(1)}_{max} + (\pi^{(2)}_{max} - \pi^{(1)}_{max}) \cdot \tau / T^{(2)}_{\pi}), \tag{5}$$

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[87]. Given the cumulative counts for first-dose-vaccinated $Z^{(1)}$ (also including the second-dose-vaccinated) and second-dose-vaccinated $Z^{(2)}$, we obtain the count for immune with vaccination as:

$$\begin{aligned} Y[t] = &\sum_{\tau=0}^{T^{(1)}_{\pi}-1} (\pi^{(1)}[\tau] \cdot (Z^{(1)}[t-\tau] - Z^{(1)}[t-\tau-1]) + \pi^{(1)}_{max} \cdot Z^{(1)}[t - T^{(1)}_{\pi}] + \\ &\sum_{\tau=0}^{T^{(2)}_{\pi}-1} ((\pi^{(2)}[\tau] - \pi^{(1)}_{max}) \cdot (Z^{(2)}[t-\tau] - Z^{(2)}[t-\tau-1]) + (\pi^{(2)} - \pi^{(1)}_{max}) \cdot Z^{(2)}[t - T^{(2)}_{\pi}] - L[t-1], \end{aligned} \tag{6}$$

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

$$L[t] = L[t-1] + Y[t]/T_L, \tag{7}$$

where T_L denotes the time-scale for losing immunity. We use $T_L = 180$ days[87]. 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 as in standard compartmental models, 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 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 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[88] for each variable v_i from Table 2 based on additional *features* $\mathrm{cov}(v_i, t)$ at different time t. The features we consider include (i) the set of static features $\mathscr{S}$, such as population density, and (ii) $\{f[t-j]\}_{f \in \mathscr{F}_i, j=1,\ldots,k}$ the set of time-varying features (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

$$v_i[t] = v_{i,L} + (v_{i,U} - v_{i,L}) \cdot \sigma\left(c + b_i + \mathbf{w}^\top \mathrm{cov}(v_i, t)\right), \tag{8}$$

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 $[v_{i,L}, v_{i,U}]$[3], which is important to stabilize training and avoid overfitting. We note that although Eq. (8) denotes a linear decomposition for $v_i[t]$ at each timestep, the overall behavior is still highly non-linear due to the relationships between compartments.

Feature forecasting: The challenge of using Eq. (8) 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 use XGBoost[89] 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.

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. (8) 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 $L_{ls} = \lambda_{ls} \sum_i |b_i|^2$ 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,

[3] 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.

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 availability of the observations $Y[T_s : T]$[4], for $Y \in \{Q, H, C, V, D, R^{(d)}\}$ [5], and consider forecasting the next τ days, $\hat{Y}[T+1 : T+\tau]$.

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 $[T_s, T_e]$:

$$L_{fit}[T_s : T] = \sum_{Y \in \{Q,H,C,V,D,R^{(d)}\}} \lambda_Y \sum_{t=T_s}^{T-\tau} \sum_{i=1}^{\tau} \frac{\mathbb{I}(Y[t+i])}{\sum_j \mathbb{I}(Y[j]) \cdot Y[j]} \cdot q(t+i-T_s; z) \cdot L(Y[t+i], \hat{Y}[t+i]). \tag{9}$$

$\mathbb{I}(\cdot) \in \{0,1\}$ 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 λ_Y are the importance 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_e (see Eq. (3)). There are rich literature in epidemiology on R_e to give us good priors on the range of the number should be. For a reproduction number $R_e[t]$ at time t, we consider the regularization

$$L_{R_e}[T_s : T] = \textstyle\sum_{t=T_s}^{T} \exp\left((R_e[t] - R)_+\right),$$

where R is a prespecified *soft* upper bound. The regularization favors the model with R_e in a reasonable range in addition to good absolute forecasting numbers. In the experiment, we set $R = 5$ without further tuning. 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

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

Last, ignoring the perturbation of a small local window, the trend of 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 encourage $v_Y[t] \approx v_Y[t-1]$, which causes linear forecasting, and cannot capture the rapid growing cases. Instead, we relax the smoothness to be on the second-order difference, *acceleration*, which is defined as $a_Y[t] = v_Y[t] - v_Y[t-1]$.

[4] We use the notation $S_i[T_s : T]$ to denote all timesteps between T_s (inclusive) and T (inclusive).

[5] Here, we denote them to represent the values for all locations, i.e. they are L-dimensional.

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

$$\mathscr{L}(T_s,T) = L_{fit}[T_s:T] + \lambda_{ls}\cdot L_{ls} + \lambda_{R_e}\cdot L_{R_e}[T_s:T] + \lambda_{dir}\cdot L_{dir} + \lambda_{acc}\cdot L_{acc}[T_s:T], \tag{10}$$

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

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[51] 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[90], 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 pseudo code 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 than 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 θ, 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[91] 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. Our framework allows the capability of modeling prediction interval forecasts, for which we replacing the L2 loss with weighted interval loss (WIS)[24] in Eq. (9) 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[24].

Algorithm 1 Pseudo-code 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 $\lambda_{R_e}$ and $\lambda_{ls}$.
Initialize trainable parameters $\theta = \{\mathbf{w_i},\ c,\ b_i\}$, and initial conditions for the compartments $\hat{E}[0]$, $\hat{I}^{(d)}[0]$, $\hat{I}^{(u)}[0]$, $\hat{R}^{(d)}[0]$, $\hat{R}^{(u)}[0]$, $\hat{H}[0]$, $\hat{C}[0]$, $\hat{V}[0]$, $\hat{D}[0]$
Split $\tau$ day validation $Y_i[T-\tau:T]$ for all locations $i$, where $Y \in \{Q,H,C,V,D,R^{(d)}\}$
while until convergence do
    Sample initial conditions $E_i[0]$, $I_i^{(d)}[0]$, $I_i^{(u)}[0]$, $R_i^{(d)}[0]$, $R_i^{(u)}[0]$, $H_i[0]$, $C_i[0]$, $V_i[0]$, $D_i[0]$
    Sample random training range $T_r$ from $[T_s, T-\tau-1]$
    $\theta \leftarrow \theta - \text{RMSProp}(\nabla_\theta \mathscr{L}(T_s, T_r))$
    Update the optimal parameters: $\theta_{opt} = \theta$ if $L_{fit}[T-\tau:T]$ is the current-best
Final fine-tuning: fine-tune with joint training and validation data:
$\theta \leftarrow \theta_{opt}$
for $F$ iterations do
    $\theta \leftarrow \theta - \text{RMSProp}(\nabla_\theta \mathscr{L}(T_s, T))$
    Update the optimal parameters: $\theta_{opt} = \theta$ if $L_{fit}[T-\tau:T]$ is the currently best
Output: Return $\theta_{opt}$
```

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 Materials.

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 computes 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 like systematic under-prediction or over-prediction. 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 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 Diebold-Mariano (DM) test[92] is used to compare our models' forecasts to those of other comparison models from "covid19-forecast-hub" (https://covid19forecasthub.org/). The four week ahead forecasts are compared using MAE and MAPE values after they had been averaged across all the locations for the dates when for both of the models produced forecasts. The p-values from the tests are adjusted using the Holm–Bonferroni method[93] to account for the multiple comparisons and KPSS tests[94] 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[95–99] as well as how they may affect healthcare access[100]. 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[101] 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[102–106]. 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. For each day in the retrospective period, we train an ensemble of $k = 5$ models to use their disagreement as a metric of uncertainity[102]. Producing 28-day forecasts from each model, we consider the two values for each location: (1) the metric performance of the single best model over the 28-day forecast (in MAE or MAPE), and (2) the variance in predictions across the k models for each day, averaged over the 28-day period. For each location, we then collect the set of average predicted variances for all release dates and compute ten quantiles at the $[10\%, 20\%, \ldots, 90\%, 100\%]$ 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 results in average performance at ten quantiles of uncertainty values, which we visualize in the form of a rejection diagram[107,108] (Extended Data Figure 1). We find that on average, the reliability of our framework can be improved through the proposed method of model uncertainty quantification by providing more caution signals to the human users on the the days of lower confidence (Extended Data Figure 1, Supplementary Section: Uncertainty Analysis).

Reporting summary

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

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). Our experimental framework relies on proprietary libraries and we are unable to publicly release this code. We detail all the experiments and implementation details in the methods section and in the supplementary figures to allow for independent replication of all the results.

Data availability

The data used for the training, validation and test sets is publicly available. All data was collected entirely from openly available sources. The access information for all sources is provided in the methods section. The dashboard showing our forecasts can be accessed from https://g.co/covidforecast.

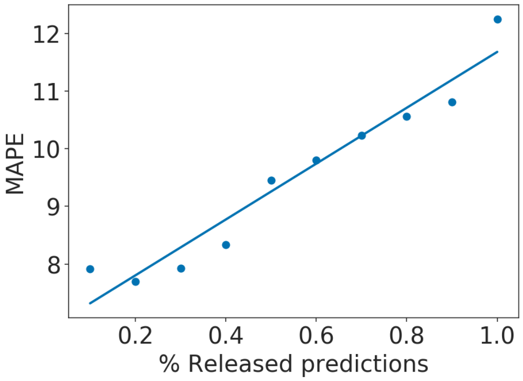
Competing interests

H.M. and S.N. are recipients of a Google.org Fellowship grant. The authors have no other competing interests to disclose.

Correspondence and requests for materials should be addressed to soarik@google.com.

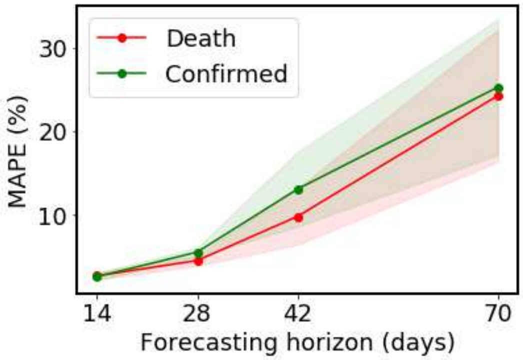
Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
AAPE	Aggregate Average Percentage Error
AE	Absolute Error
APE	Average Percentage Error
BMI	Body Mass Index
CI	Confidence Interval
CLI	COVID like illness
COVID-19	Coronavirus Disease 2019
DFE	Disease-Free Equilibrium
EPA	Environmental Protection Agency
$F^{(u)}$	Force of Infection of Undocumented
GDP	Gross Domestic Product
HIPAA	Health Insurance Portability and Accountability Act
ICU	Intensive Care Unit
MAE	Mean Average Error
MAPE	Mean Average Percentage Error
NMAE	Normalized Mean Average Error
NPI	Non-Pharmaceutical Intervention
PPE	Personal Protective Equipment
R_{eff}	Effective Reproductive Number
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SEIR	Susceptible, Exposed, Infectious, Recovered Compartment Model
SIR	Susceptible, Infectious, Recovered Compartment Model
SOE	State of Emergency in Japan
USA	United States of America
WIS	Weighted Interval Loss Score

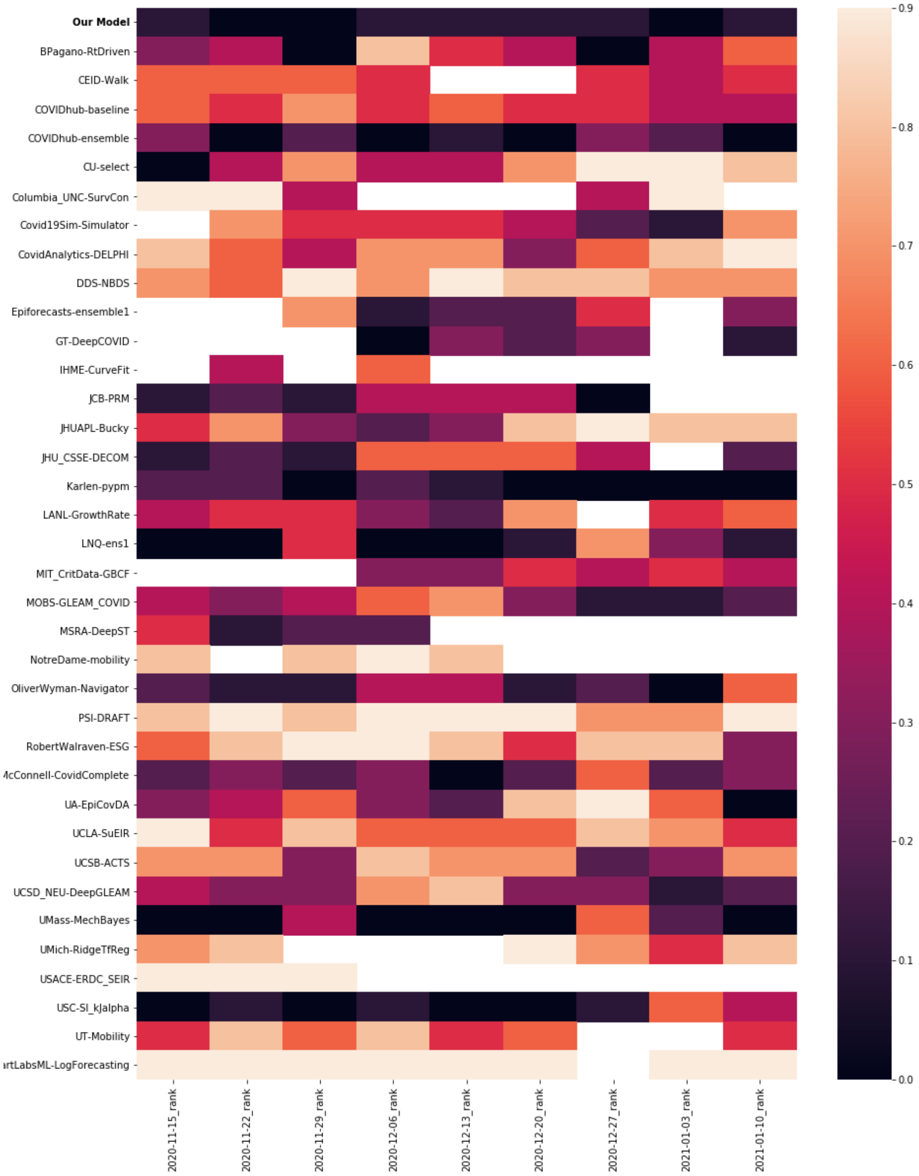


Extended Data Figure 1. A rejection diagram showing the percentage of dates on which a prediction is made, after thresholding on model disagreement due to model uncertainty, versus the MAPE performance on those dates. From this, we can see that better average metric performance (on the days for which 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, $R^2 = 0.941, f(x) = 2.18x + 9.50$.

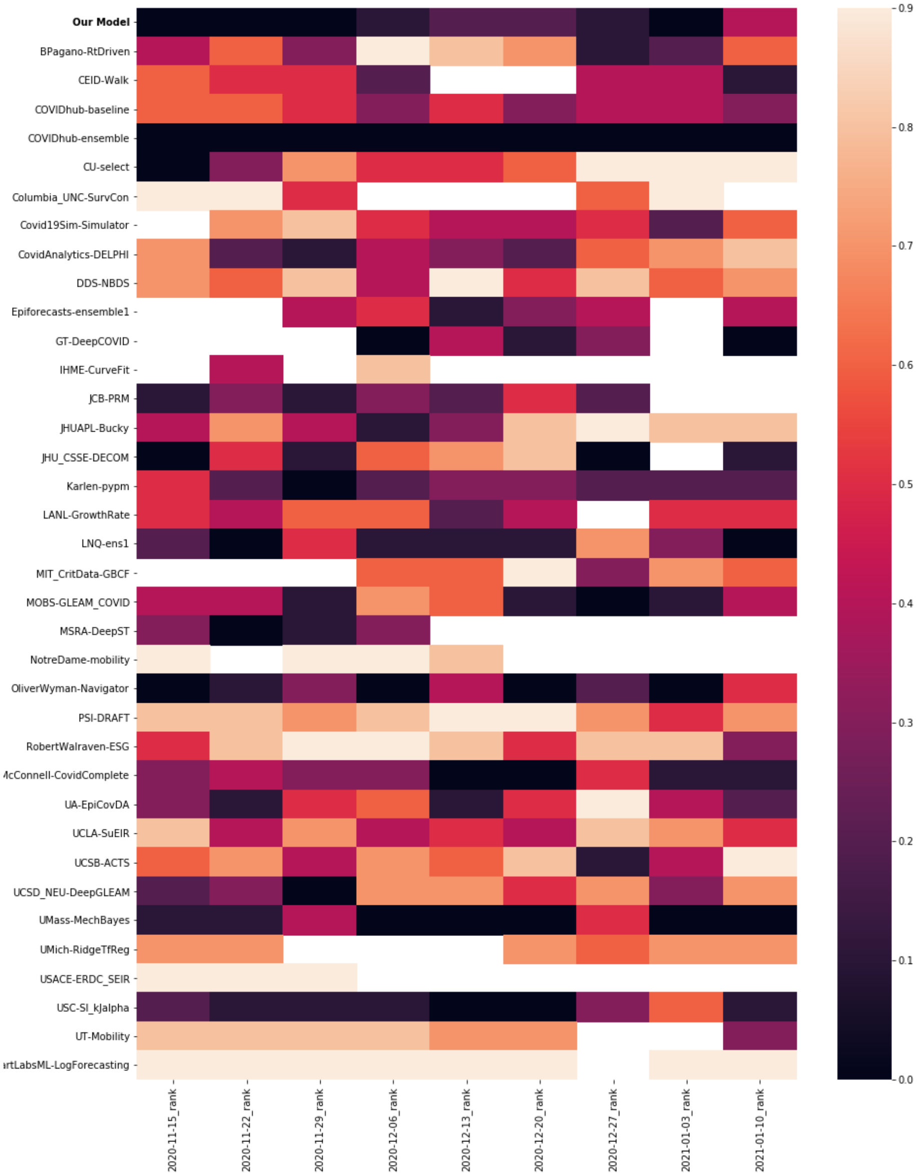
677 # Extended Data Figures and Tables



Extended Data Figure 2. MAPE (%) vs. forecasting horizon, averaged across different prediction dates with the 95 % confidence intervals (shaded). For each forecasting horizon, we retrain the model after changing the corresponding τ value in Algorithm 1. Forecasting different horizons may improve utility, as having insights farther into the future can allow better policy and health planning. On the other hand, as there is increasing uncertainty into the future and the value of observed features decay over time, it is expected that the model accuracy should get worse. We choose 28-day horizon considering the trade-off between the accuracy and utility. For forecast values beyond 4 weeks, we often observe significant degradation in accuracy.



Extended Data Figure 3. Normalized ranking for death MAE in the prospective evaluation period. The darker the color, the higher the ranking of the model is for the corresponding prediction date.



Extended Data Figure 4. Normalized ranking for death MAPE in the prospective evaluation period. The darker the color, the higher the ranking of the model is for the corresponding prediction date.

Extended Data Table 1. Results from a fairness analysis of forecast performance. The partial correlations between variables of interest and the MAPE for USA counties after conditioning on the other demographic variables. The correlations are calculated using Kendall's Tau and the impact of the other variables is controlled for using linear regression.

USA Model			
Variable	Tau	p-value	Adjusted p-value
Population Density	-0.152	< 0.001	< 0.001
Median Age	0.072	< 0.001	< 0.001
Median Income	-0.059	< 0.001	< 0.001
Fraction Female	-0.041	0.001	0.003
Fraction White	-0.025	0.037	0.112
Fraction Hispanic	-0.024	0.047	0.112
Fraction Black	-0.010	0.397	0.397
Japan Model			
Variable	Tau	p-value	Adjusted p-value
Percentage Female	-0.244	0.017	0.149
Population Density	0.235	0.021	0.171
Korean	0.154	0.132	0.925
Japanese	0.123	0.229	1.000
Chinese	-0.074	0.466	1.000
Older	-0.063	0.538	1.000
Median Income	0.061	0.551	1.000
Middle	-0.061	0.551	1.000
Younger	-0.057	0.576	1.000

Extended Data Table 2. Model feature importance. Top 10 time-series and top 5 static feature importance ranks for the average undocumented contact rate in the USA and Japan models.

USA Model		Japan Model	
Time series features		Time series features	
Feature name	Median rank	Feature name	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		Static features	
Ratio of 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/100k Population	4
Number of Households	5	Number of Doctors/100k Population	5

Extended Data Table 3. Counterfactual analysis for combined mobility restrictions in Japan In a combined scenario, we decrease a certain percentage (strong: 90%, medium: 60%, weak: 30%) of the mobility for Park, Work, Transit, Grocery, and Retail together. We also increase a certain percentage (strong: 300%, medium: 200%, weak; 100%) of the mobility for residential, to reflect the altered behavioral patterns associated with mobility restrictions. The numbers represent the increase in confirmed cases predicted to occur over the next 4 weeks.

	Strong	Medium	Weak
Overall	-97115 (-16.6%)	-73997 (-12.8%)	-42115 (-7.4%)
Tokyo	-27514 (-24.4%)	-20947 (-18.6%)	-11893 (-10.6%)
Chiba	-22453 (-43.1%)	-16975 (-32.6%)	-9511 (-18.3%)
Kanagawa	-10210 (-24.0%)	-7859 (-18.5%)	-4543 (-10.7%)
Fukuoka	-10953 (-37.3%)	-8249 (-28.1%)	-4619 (-15.7%)
Osaka	-5976 (-19.2%)	-4550 (-14.6%)	-2599 (-8.4%)
Saitama	-4204 (-22.1%)	-3248 (-17.1%)	-1890 (-10.0%)
Hyogo	-2901 (-24.3%)	-2248 (-18.8%)	-1316 (-11.0%)
Kyoto	-3011 (-30.4%)	-2265 (-22.9%)	-1268 (-12.8%)

Extended Data Table 4. Counterfactual analysis for individual mobility restrictions in Japan The change in case numbers associated with instituting strong measures to reduce mobility in Japan. We decrease 90% of the mobility to areas allocated to Park, Work, Transit, Grocery, and Retail in the mobility data separately. We also separately increase the mobility to residential by 300% to reflect the altered behavioral patterns associated with mobility restrictions. The numbers represent the increase in confirmed cases predicted to occur over the next 4 weeks.

	Park	Work	Transit	Grocery	Retail	Residential
Overall	-5443 (-1.0%)	-18943 (-3.4%)	-8013 (-1.5%)	-18894 (-3.5%)	-20833 (-3.7%)	-56311 (-9.7%)
Tokyo	-1610 (-1.4%)	-5328 (-4.7%)	-2215 (-2.0%)	-5195 (-4.6%)	-5859 (-5.2%)	-16018 (-14.2%)
Chiba	-1098 (-2.1%)	-4186 (-8.0%)	-1750 (-3.4%)	-4093 (-7.9%)	-4598 (-8.8%)	-13076 (25.1%)
Kanagawa	-606 (-1.4%)	-2116 (-5.0%)	-890 (-2.1%)	-2090 (-4.9%)	-2329 (-5.5%)	-5927 (-13.9%)
Fukuoka	-606 (2.1%)	-2057 (-7.0%)	-893 (-3.0%)	-2127 (7.2%)	-2281 (-7.8%)	-6097 (-20.8%)
Osaka	-329 (-1.1%)	-1202 (-3.9%)	-487 (-1.6%)	-1204 (-3.9%)	-1282 (-4.1%)	-3440 (-11.1%)
Saitama	-245 (-1.3%)	-834 (-4.4%)	-379 (-2.0%)	-904 (-4.8%)	-970 (-5.1%)	-2475 (-13.0%)
Hyogo	-170 (-1.4%)	-595 (-5.0%)	-262 (-2.2%)	-612 (-5.1%)	-685 (-5.7%)	-1724 (-14.4%)
Kyoto	-165 (-1.7%)	-576 (-5.8%)	-241 (-2.4%)	-583 (-5.9%)	-609 (-6.2%)	-1687 (-17.0%)

Extended Data Table 5. Counterfactual analysis for applying NPIs in tandem with vaccinations drives in USA states. The change is the predicted number of Susceptible Individuals on the 28th day of the forecasting horizon for the Top-5 states by baseline susceptible individuals. NPI levels are from Rand Corporation[109]. Negative percentages imply a reduction in counts, positive percentages imply an increase. See Supplementary Materials for detailed results and discussion.

Vaccination Drives	Predicted features Baseline	Rand Level 1	Rand Level 3	Rand Level 5
Baseline	Illinois: 6122387 California: 6073129 Florida: 4732335 Pennsylvania: 4520196 Georgia: 3796765	Illinois: 5033620 (-17.78%) California: 5260227 (-13.39%) Florida: 6635598 (40.22%) Pennsylvania: 4785996 (5.88%) Georgia: 4452852 (17.28%)	Illinois: 6038620 (-1.37%) California: 7031898 (15.79%) Florida: 8119648 (71.58%) Pennsylvania: 5303852 (17.34%) Georgia: 5027352 (32.41%)	Illinois: 6319586 (3.22%) California: 7467760 (22.96%) Florida: 8552248 (80.72%) Pennsylvania: 5392968 (19.31%) Georgia: 5144156 (35.49%)
0.5% pop. vaccinated/day	Illinois: 4593573 (-24.97%) California: 1710801 (-71.83%) Florida: 2814891 (-40.52%) Pennsylvania: 3083091 (-31.79%) Georgia: 2688499 (-29.19%)	Illinois: 3657703 (-40.26%) California: 1181140 (-80.55%) Florida: 4303318 (-9.07%) Pennsylvania: 3300274 (-26.99%) Georgia: 3234374 (-14.81%)	Illinois: 4520088 (-26.17%) California: 2311234 (-61.94%) Florida: 5541122 (17.09%) Pennsylvania: 3740164 (-17.26%) Georgia: 3732658 (-1.69%)	Illinois: 4766381 (-22.15%) California: 2626504 (-56.75%) Florida: 5918076 (25.06%) Pennsylvania: 3819714 (-15.5%) Georgia: 3837885 (1.08%)

Extended Data Table 6. Counterfactual analysis on predicted cases while applying mobility restrictions alongside vaccinations in Japan. The change in predicted confirmed cases for five prefectures across different scenarios. The scenario with no mobility restrictions is normalized to a baseline where no overrides are applied. The scenarios with mobility restriction are referenced to a baseline where just a vaccination drive is applied. Low and high vaccination rate scenarios are modelled, representing 0.1% and 1% of the population vaccinated daily respectively and in both scenarios assuming a 95% effectiveness.

Vaccination Scenario	NPI Scenario	Tokyo	Kanagawa	Osaka	Saitama	Chiba
Forecasted Features Baseline	Forecasted Features Baseline	281847	79820	74888	43358	38199
0.1% pop. vaccinated/day	Forecasted Features Baseline	280481 (-0.49%)	79461 (-0.45%)	74862 (-0.03%)	43273 (-0.2%)	38108 (-0.24%)
0.1% pop. vaccinated/day	Weak Mobility Restrictions	263068 (-6.66%)	74791 (-6.3%)	74482 (-0.54%)	42098 (-2.91%)	36900 (-3.4%)
0.1% pop. vaccinated/day	Medium Mobility Restrictions	258057 (-8.44%)	73465 (-7.96%)	74366 (-0.7%)	41752 (-3.7%)	36552 (-4.31%)
0.1% pop. vaccinated/day	Strong Mobility Restrictions	255473 (-9.36%)	72777 (-8.82%)	74301 (-0.78%)	41568 (-4.13%)	36370 (-4.79%)
1% pop. vaccinated/day	Forecasted Features Baseline	276428 (-1.92%)	78400 (-1.78%)	74785 (-0.14%)	43021 (-0.78%)	37837 (-0.95%)
1% pop. vaccinated/day	Weak Mobility Restrictions	261728 (-7.14%)	74457 (-6.72%)	74455 (-0.58%)	42015 (-3.1%)	36812 (-3.63%)
1% pop. vaccinated/day	Medium Mobility Restrictions	257416 (-8.67%)	73308 (-8.16%)	74352 (-0.71%)	41712 (-3.8%)	36510 (-4.42%)
1% pop. vaccinated/day	Strong Mobility Restrictions	255160 (-9.47%)	72701 (-8.92%)	74294 (-0.79%)	41548 (-4.17%)	36350 (-4.84%)

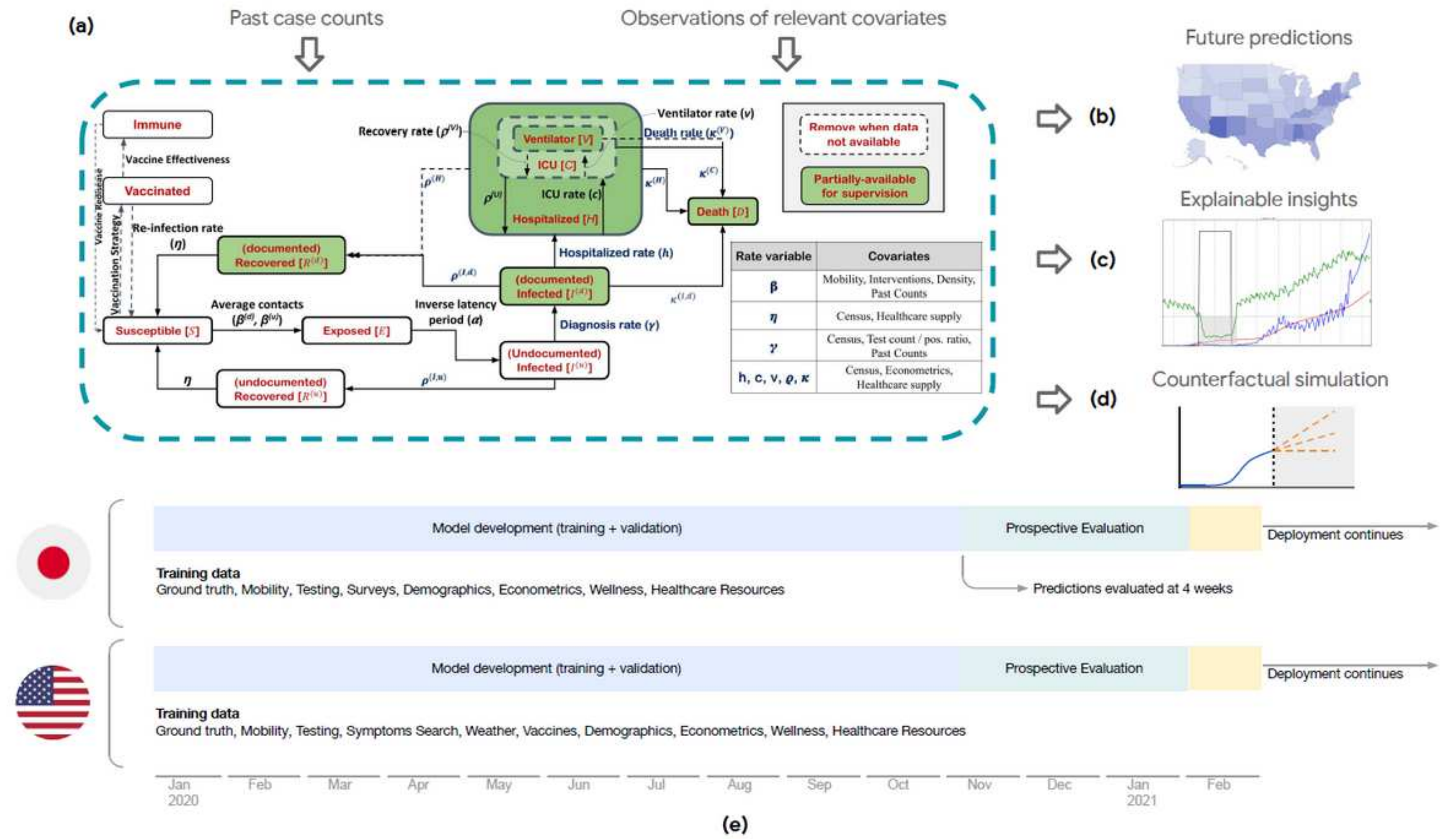
Extended Data Table 7. Counterfactual analysis on predicted deaths while applying mobility restrictions alongside vaccinations in Japan. The change in predicted COVID-19 associated deaths for five Japanese prefectures across different scenarios. The scenario with no mobility restrictions is normalized to a baseline where no overrides are applied. The scenarios with mobility restriction are referenced to a baseline where just a vaccination drive is applied. Low and high vaccination rate scenarios are modelled, representing 0.1% and 1% of the population vaccinated daily respectively and in both scenarios assuming a 95% effectiveness.

Vaccination Scenario	NPI Scenario	Tokyo	Kanagawa	Osaka	Saitama	Chiba
Forecasted Features Baseline	Forecasted Features Baseline	2920	837	1689	582	417
0.1% pop. vaccinated/day	Forecasted Features Baseline	2908 (-0.41%)	835 (-0.24%)	1689 (-0.0%)	581 (-0.0%)	417 (-0.0%)
0.1% pop. vaccinated/day	Weak Mobility Restrictions	2737 (-6.27%)	810 (-3.23%)	1683 (-0.3%)	575 (-1.2%)	412 (-1.2%)
0.1% pop. vaccinated/day	Medium Mobility Restrictions	2687 (-7.98%)	802 (-4.18%)	1682 (-0.41%)	573 (-1.55%)	411 (-1.68%)
0.1% pop. vaccinated/day	Strong Mobility Restrictions	2661 (-8.87%)	798 (-4.66%)	1681 (-0.47%)	571 (-1.72%)	410 (-1.92%)
1% pop. vaccinated/day	Forecasted Features Baseline	2874 (-1.58%)	830 (-0.84%)	1688 (-0.06%)	580 (-0.34%)	416 (-0.24%)
1% pop. vaccinated/day	Weak Mobility Restrictions	2726 (-6.64%)	808 (-3.46%)	1683 (-0.36%)	574 (-1.2%)	412 (-1.44%)
1% pop. vaccinated/day	Medium Mobility Restrictions	2682 (-8.15%)	801 (-4.18%)	1682 (-0.41%)	572 (-1.55%)	410 (-1.68%)
1% pop. vaccinated/day	Strong Mobility Restrictions	2658 (-8.94%)	798 (-4.66%)	1681 (-0.47%)	571 (-1.72%)	410 (-1.92%)

Extended Data Table 8. Counterfactual analysis on the predicted consequences of delays in applying NPIs. A comparison of predicted confirmed case reductions for NPIs applied immediately compared with if a delay of seven days is introduced. National forecasts are shown alongside those for the prefectures with the most confirmed cases. The ’Combined Medium’ scenario is a 60% decrease in the Google Mobility values for Park, Work, Transit, Grocery, and Retail mobility and a 20% increase in Google Mobility values for Residential mobility. For the ’Combined Strong’ scenario, the corresponding changes are a 90% decrease and a 30% increase, respectively.

Location	Combined Medium	Combined Medium Delayed 7-days	1 Combined Strong	Combined Strong Delayed 7-days
All Japan	-36521	-20203 (+45.0%)	-41829	-22891 (+45.2%)
Hokkaido	-279	-157 (+44.0%)	323	181 (+44.0%)
Saitama	-2072	-1118 (+46.0%)	-2375	-1274 (+46.4%)
Osaka	-908	-506 (+44.2%)	-1033	-579 (+44.0%)
Tokyo	-23066	-12990 +(43.6%)	-26157	-14569 (+44.3%)
Gunma	-59	-36 (+39.0%)	-69	-42 (+39.0%)

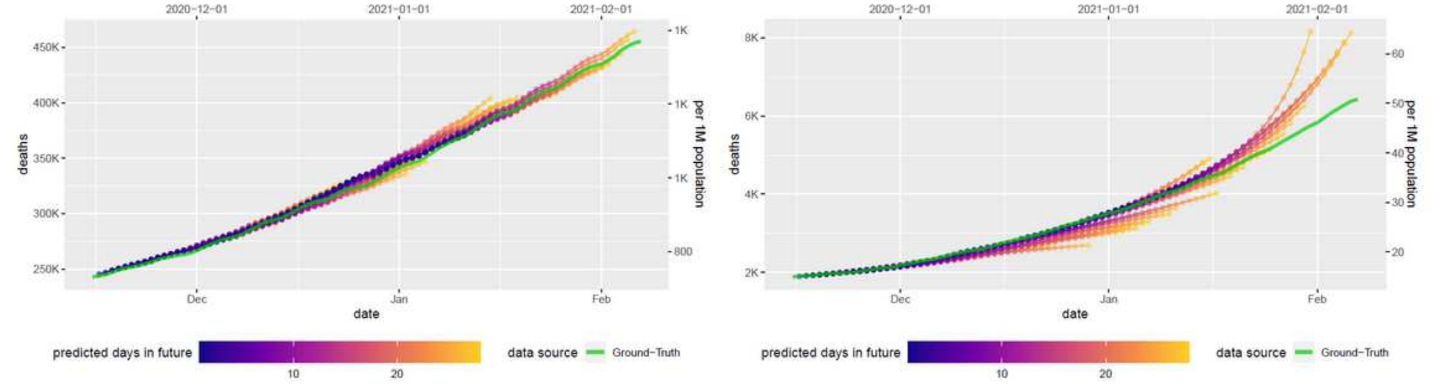
Figures


(a)
Past case counts
Observations of relevant covariates
Immune
Vaccine Effectiveness
Vaccinated
Vaccine Redisease
Vaccination Strategy
Re-infection rate
(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 [$I^{(d)}$]
$\rho^{(I,d)}$
$\kappa^{(I,d)}$
Hospitalized rate (h)
Recovery rate ($\rho^{(V)}$)
$\rho^{(H)}$
$\rho^{(C)}$
Ventilator [V]
ICU [C]
ICU rate (c)
Hospitalized [H]
Ventilator rate (v)
Death rate ($\kappa^{(V)}$)
$\kappa^{(C)}$
$\kappa^{(H)}$
Death [D]
Remove when data not available
Partially-available for supervision
(b)
(c)
(d)
Future predictions
Explainable insights
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)

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

Figure 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) model22, 23. 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 Japan at the prefecture level. (c) Interpretable elements, including predictions for the effective R number and force of infection provide explainable and actionable insights. (d) Simulations of counterfactual scenarios can be used to estimate the 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.

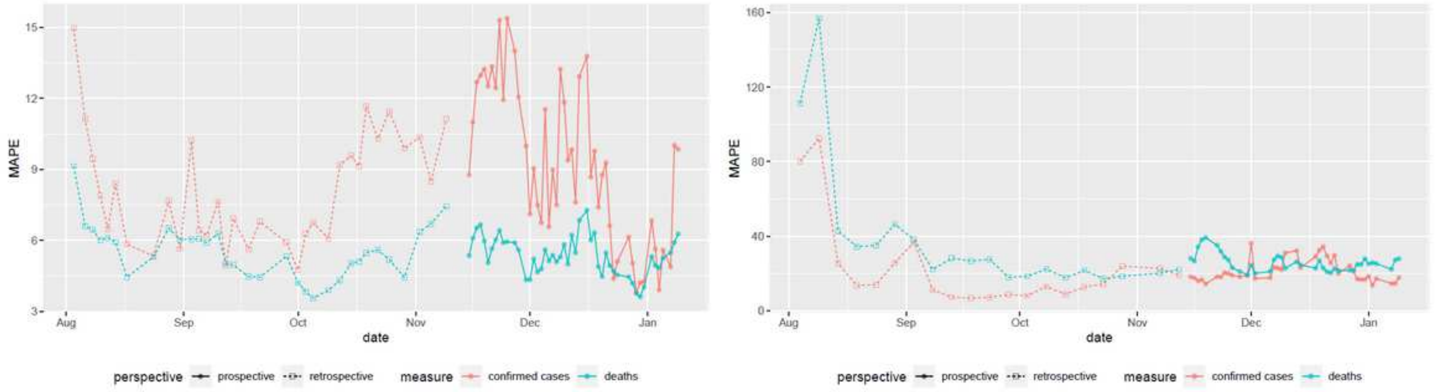


(a)

(b)

Figure 2

Prospective forecasts for the USA 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.

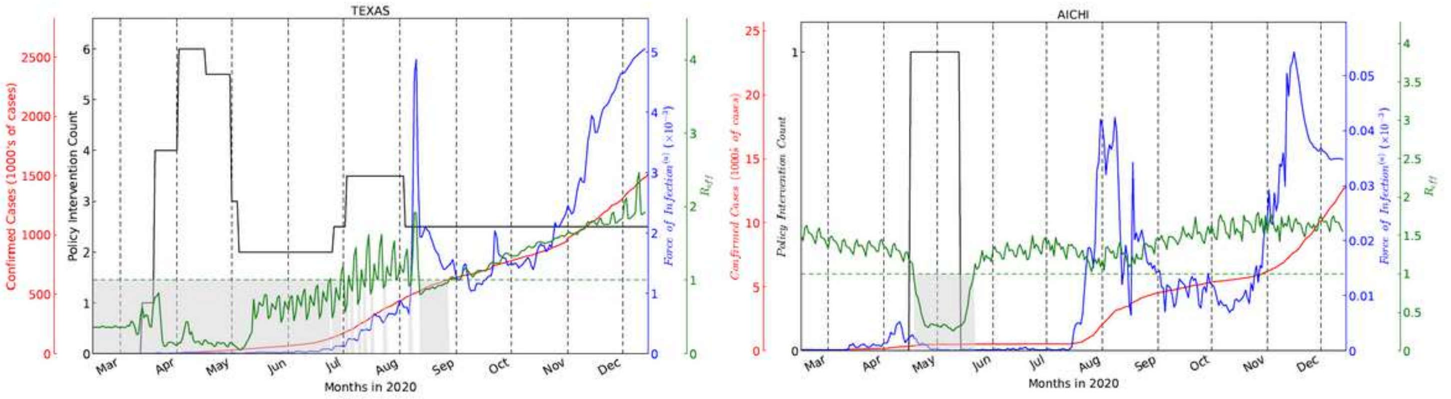


(a)

(b)

Figure 3

Retrospective and prospective 28-day MAPE over time. Performance over time is shown for the (a) state-level USA models (b) prefecture-level Japan model. 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).

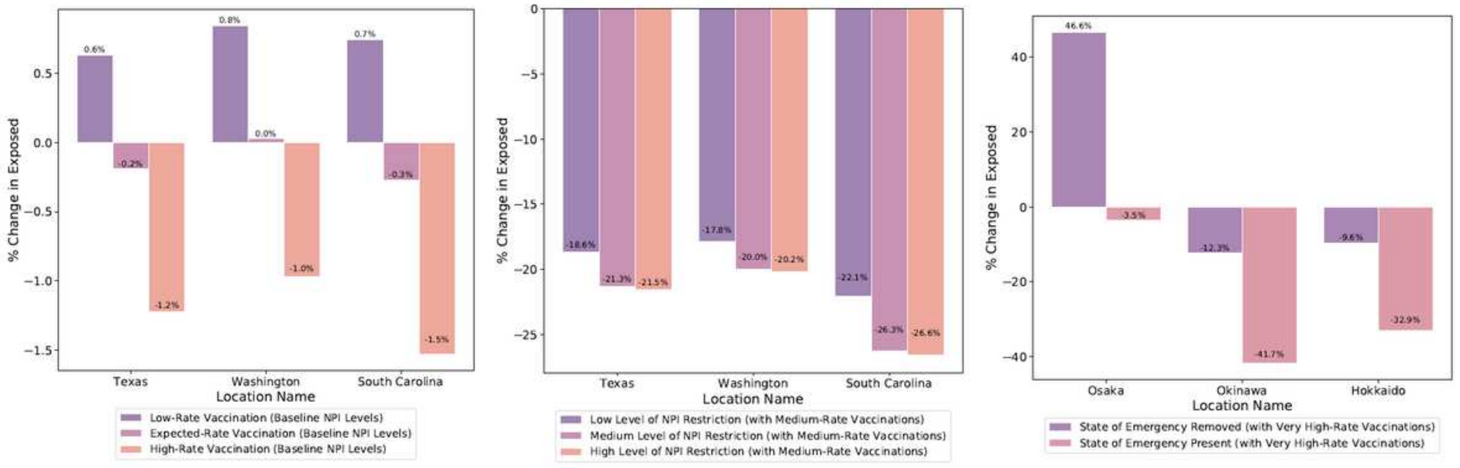


(a)

(b)

Figure 4

Interpretable model outputs. Confirmed cases, number of NPIs, F(u) and Reff for Texas, USA (a) and Aichi, Japan (b), chosen to represent a location with high and low numbers of COVID-19 associated deaths respectively. Confirmed case counts and number of Non-Pharmaceutical Interventions (NPIs) are plotted on the left Y-axis, and F(u) (see Eq. 1) and the Reff (see Eq. 3) are plotted on the right Y-axis. For Reff < 1 (shaded grey regions below the horizontal dotted line), dynamics are tending towards the Disease-Free Equilibrium (DFE)31. These areas often overlap with the dates when multiple NPIs are imposed.



(a)

(b)

(c)

Figure 5

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 baseline exposed individual counts are 69694, 67591 and 63742 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, more than 17

%across the three states. 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 very high rate vaccination (2% population/day), and considering the cases of applying or removing the State of Emergency. Here, the baseline exposed individual counts are 5779, 3838 and 3253 for Osaka, Okinawa and Hokkaido respectively. Applying the the State of Emergency is observed to be highly 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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ariksupplementary.pdf